Анализ таблиц сопряженности

Малов Сергей Васильевич

Санкт-Петербургский государственный электротехнический университет

12/19 декабря 2020 г.

План

Вероятностная модель эксперимента

Наблюдение - дискретный случайный вектор $T = (T_1, \dots, T_r)$ с конечным числом возможных значений для каждой компоненты (признака)

- Возможные значения T_j – уровни j-го признака
- Величины $T_1, \dots, T_r$ могут быть качественными или ординальными
- Не умаляя общности считаем $T_j \in \{1, \dots, d_j\}$, $j = 1, \dots, r$
  - для ординальных признаков порядок уровней сохраняется
- совместное распределение задается вероятностями $p_{i_1 \dots i_r} = \Pr(T_1 = i_1, \dots, T_r = i_r)$
  - $p_{i_1 \dots i_r} \geq 0, \ \forall \ i_1, \dots, i_r$
  - $\sum_{i_1 \dots i_r} p_{i_1 \dots i_r} = 1$
- Модель статистического эксперимента - параметрическая
  - параметр распределения: $p_{i_1, \dots, i_r} \in \Theta$

$$\Theta = \{p_{i_1, \dots, i_r} : p_{i_1, \dots, i_r} \geq 0, \sum_{i_1 \dots i_r} p_{i_1 \dots i_r} = 1\}$$

  - размерность параметра: $i_1 \cdot \ldots \cdot i_r - 1$

Получение и редукция данных

Исходные данные - выборка $(T_1, \ldots, T_n)$ из распределения T: $T_s = (T_{s1}, \ldots, T_{sr})$, $s = 1, \ldots, n$

- Функция правдоподобия $L(T; p) = \prod_{i_1 \ldots i_r} p_{i_1 \ldots i_r}^{n_{i_1 \ldots i_r}}$
  - $n_{i_1 \ldots i_r} = \sum_{s=1}^{n} \mathbf{1}_{\{T_{s1}=i_1, \ldots, T_{sr}=i_r\}}$ – число элементов выборки имеющих соответствующие значения компонент
  - $\sum_{i_1, \ldots, i_r} n_{i_1 \ldots i_r} = n$ – размер выборки
  - совокупность всех $n_{i_1 \ldots i_r}$ – достаточная статистика
- Значения $n_{i_1 \ldots i_r}$ образуют таблицу сопряженности признаков – массив $d_1 \times \ldots \times d_r$
- При фиксированном размере выборки $n_{i_1 \ldots i_r}$ имеют совместное мультиномиальное распределение с парметрами $p_{i_1 \ldots i_r}$ и n

План

Пуассоновский поток событий

Пусть $A_1, A_2, \ldots$ – последовательность однородных событий, происходящих в случайные моменты времени $T_1, T_2, \ldots$.

Определение: Поток событий называется простейшим (пуассоновским), если выполнены условия:

- стационарный: вероятность появления k событий в интервале $[s, s+t)$ не зависит от $s \geq 0$
- ординарный: вероятность появления двух и более событий в малом интервале времени есть величина бесконечно малая по отношению к вероятности появления одного события в этом интервале
- без последействия: числа событий, появляющихся в непересекающиеся интервалы времени, являются независимыми случайными величинами

Процесс Пуассона: точечный процесс начинающийся из нуля со скачками единичной величины в моменты появления событий простейшего потока

- интенсивность – среднее число событий, появляющихся в единицу времени

Пуассоновский поток событий

Основные свойства:

- Число событий $\nu(s,t)$ простейшего потока с интенсивностью λ, появляющихся в инетрвале времени $[s, s+t)$ имеет распределение Пуассона: $\nu(s,t) \sim \mathrm{Pois}(\lambda t)$
- Время ожидания $\tau(s)$ ближайшего события простейшего потока с интенсивностью λ, начиная с момента времени s, имеет показательное распределение: $\tau(s) \sim \mathrm{Exp}(\lambda)$
- Поток событий, получающийся слиянием двух (или более) независимых простейших потоков событий, будет простейшим с интенсивностью равной сумме интенсивностей исходных потоков событий

Мультиномиальный и пуассоновский подходы

Категориальный эксперимент

- результат эксперимента – дискретная величина с конечным числом урговней $X \in \{1, \ldots, d\}$
- параметризация: $p_i = \mathbb{P}(X = i)$, $i = 1, \ldots, d$; $(p_1, \ldots, p_d) \in \{(p_1, \ldots, p_d) : p_i \geq 0, i = 1, \ldots, d, \sum_i p_i = 1\}$
- $X_1, \ldots, X_n$ – выборка из распределения X, $n \in \mathbb{N}$
- достаточная статистика: $(n_1, \ldots, n_d)$
  - $n_s = \sum_{i=1}^n \mathbf{1}_{\{X=s\}}$ – число наблюдений, имеющих значение s, $s = 1, \ldots, d$.

Классический подход

- число экспериментов $n \in \mathbb{N}$ – фиксированное значение
- $(n_1, \ldots, n_d) \in \mathrm{Mult}(p_1, \ldots, p_d; n)$

Пуассоновский подход

- наблюдения появляются как события пуассоновского потока $A_1, A_2, \ldots$
- число наблюдений к моменту времени T – случайная величина $n \sim \text{Pois}(\lambda)$
- p_i вероятность того, что каждое наблдюдение в случае появления примет значение i
- достаточная статистика $(n_1, \ldots, n_d)$
  - $(n_1, \ldots, n_d)$ – независимые величины
  - $n_i \sim \text{Pois}(\lambda p_i)$, $i = 1, \ldots, d$

Связь пуассоновского и мультиномиального подходов

- условное распределение $(n_1, \ldots, n_d)$ при условии $\sum_{i=1}^{d} n_i = n$ – мультиномиальное $\text{Mult}(p_1, \ldots, p_d; n)$

Многомерный эксперимент, структуризация

Рассмотрим результат статистического эксперимента, включающего r признаков:
$$\boldsymbol{X} = (X_1, \ldots, X_r) : X_i \in \{1, \ldots, d_i\},\ i = 1, \ldots, r$$

- с точки зрения общей модели эксперимента
$$\boldsymbol{X} \leftrightarrow Y \in \{1, \ldots, d_1 \cdot \ldots \cdot d_r\}$$
  - для задания распределения можно использовать мультиномиальный или пуассоновский подходы
  - достаточная статистика – $\{n_{j_1 \ldots j_r}\}_{j_1 \ldots j_r}$
- для постановки задач используется многомерная структура параметра
  - большинство задач касается форм зависимости компонент совместного распределения $\boldsymbol{X}$
  - достаточную статистику удобно структурировать в массив сопряженности $\|n_{j_1 \ldots j_r}\|_{j_1 \ldots j_r}$
  - аналогично структурируется параметр распределения
    - в мультиномиальной модели $\|p_{j_1 \ldots j_r}\|_{j_1 \ldots j_r}$
    - в пуассоновской модели $\|\lambda_{j_1 \ldots j_r}\|_{j_1 \ldots j_r}$: $\lambda_{j_1 \ldots j_r} = \lambda p_{j_1 \ldots j_r}$
  - многие задачи допускают формулировку в терминах условных распределений
  - переход от совместных к условным распределениям востребован при наличии контроля некоторых признаков

План

Двумерные таблицы сопряженности

(А) Таблица сопряженности

Y	X: 1	…	d_2	Всего
1	n_{11}	…	n_{1d_2}	n_{1+}
…	…	…	…	…
i	n_{i1}	…	n_{id_2}	n_{i+}
…	…	…	…	…
d_1	n_{d_11}	…	$n_{d_1d_2}$	n_{d_1+}
Всего	n_{+1}	…	n_{+d_2}	n

(Б) Совместное распределение

Y	X: 1	…	d_2	Распр. Y
1	p_{11}	…	p_{1d_2}	p_{1+}
…	…	…	…	…
i	p_{i1}	…	p_{id_2}	p_{i+}
…	…	…	…	…
d_1	p_{d_11}	…	$p_{d_1d_2}$	p_{d_1+}
Распр. X	p_{+1}	…	p_{+d_2}	1

(В) Условное распределение

Y	X: 1	…	d_2
1	$p_{1\|1}$	…	$p_{1\|d_2}$
…	…	…	…
i	$p_{i\|1}$	…	$p_{i\|d_2}$
…	…	…	…
d_1	$p_{d_1\|1}$	…	$p_{d_1\|d_2}$
Сумма	1	1	1

- Таблица сопряженности (А) – матрица $d_1 \times d_2$
- Модель совместного распределения (Y, X)
  - элементы таблицы сопряженности имеют совместное мультиномиальное распределение с параметрами из таблицы (Б)
- Модель условного распределения Y при условии X (В)
  - считаем, что $p_{+j} > 0,\ j = 1, \dots, d_2$ и $p_{i+} > 0,\ i = 1, \dots, d_1$
  - $p_{i|j} = p_{ij}/p_{+j},\ i = 1, \dots, d_1,\ j = 1, \dots, d_2$
  - $(n_{1j}, \dots, n_{d_1j}) \in \mathrm{Mult}(p_{1|j}, \dots, p_{d_1|j}; n_{+j})$ при всех $j = 1, \dots, d_2$
- Гипотеза независимости признаков $H_0 : p_{ij} = p_{i+}p_{+j}\ \forall (i, j)$ в модели совместного распределения соответствует гипотезе однородности $H_0^* : p_{i|j} = p_{i|1}\ \forall (i, j)$ в модели условного распределения

Классические методы анализа

Проверка гипотезы независимости признаков:

- критерий χ^2
  - статистика критерия: $X^2 = \sum_{i=1}^2 \sum_{j=1}^2 \dfrac{(n_{ij} - n_{i+}n_{+j}/n)^2}{n_{i+}n_{+j}/n}$
  - асимптотическое распределение: $\chi^2_{(d_1-1)(d_2-1)}$
  - доверительная область: $X^2 \le x_\alpha,\ x_\alpha : K_{(d_1-1)(d_2-1)}(x_\alpha) = 1 - \alpha$
  - P-значение: $pv = 1 - K_{(d_1-1)(d_2-1)}(X^2)$
- критерий отношения правдоподобия
  - статистика критерия: $G^2 = 2\sum_{i=1}^2 \sum_{j=1}^2 n_{ij} \log(n_{ij}n/(n_{i+}n_{+j}))$
  - асимптотическое распределение: $\chi^2_{(d_1-1)(d_2-1)}$
  - доверительная область: $X^2 \le x_\alpha,\ x_\alpha : K_{(d_1-1)(d_2-1)}(x_\alpha) = 1 - \alpha$
  - P-значение: $pv = 1 - K_{(d_1-1)(d_2-1)}(G^2)$
- критерий на линейность зависимости (Trend test)
  - статистика критерия: $M^2 = (n-1)r^2$ (r^2 – коэффициент корреляции Пирсона)
  - асимптотическое распределение: χ^2_1
  - доверительная область: $M^2 \le x_\alpha,\ x_\alpha : K_1(x_\alpha) = 1 - \alpha$
  - P-значение: $pv = 1 - K_1(G^2)$

Сопряженность двух признаков

Модель логистической регрессии

- применима для анализа таблиц сопряженности признаков (Y, X) размера $2 \times d$
- для удобства используем уровни признака $Y \in \{0, 1\}$
- введем условные вероятности $\pi_i = p_{1|j}$, $j = 1, \ldots, d$
- допускается контроль признака X (независимая переменная)
- условные распределения – Бернулли
- обобщенная линейная модель

$$\mathrm{logit}(\pi_i) = \alpha + \beta_i, \ i = 1, \ldots, d$$

  - правая часть – однофакторный дисперсионный анализ
  - α – взвешенное среднее
  - β_i – главные эффекты: $\beta_* = 0$
- гипотеза независимости $H_0 : \beta_1 = \ldots = \beta_d = 0$
- $\beta_i - \beta_j$ – логарифм частного отношения шансов

Сопряженность двух признаков

Проверка гипотезы независимости $H_0 : \beta_1 = \ldots = \beta_d = 0$

- Критерий типа Вальда
  - $H_0 : \boldsymbol{\psi} = 0$, $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_{d-1})'$ – линейно независимые сравнения
    - например, $\psi_i = \beta_i$, если веса $v_i < 1$, $i = 1, \ldots, d-1$
  - статистика критерия: $W = \hat{\boldsymbol{\psi}}' \hat{\Gamma}_\psi^{-1} \hat{\boldsymbol{\psi}}$
    - $\hat{\Gamma}_\psi^{-1}$ – оценка матрицы ковариации сравнений $\boldsymbol{\psi}$
    - значение W не зависит от выбора сравнений
  - предельное распределение: χ^2_{d-1}
  - P-значение: pv=$1 - K_{d-1}(W)$
- Критерий отношения правдоподобия
  - статистика критерия: $G = 2(LL_S - LL_A)$ (deviance)
    - LL_S максимум логарифма правдоподобия в общей модели
    - LL_A максимум логарифма правдоподобия в аддитивной модели
  - предельное распределение: χ^2_{d-1}
  - P-значение: pv=$1 - K_{d-1}(G)$
- в отличие от классической модели $W \neq G$ в общем случае

Сопряженность двух признаков

Лог-линейная модель

- Применима для анализа таблиц сопряженности признаков (Y, X) размера $d_1 \times d_2$ (общего вида)
- Параметры распределения $\lambda_{ij} = \lambda p_{ij}$, $i = 1, \ldots, d_1$, $j = 1, \ldots, d_2$
- Пуассоновская схема
  - n_{ij} независимы
  - n_{ij} имеют пуассоновское распределение $\mathrm{Pois}(\lambda_{ij})$
- Обобщенная линейная модель
$$\log(\lambda_{ij}) = \alpha + \beta_i^Y + \beta_j^X + \beta_{ij}^{YX},$$
  - правая часть – двуфакторный дисперсионный анализ
  - α – взвешенное среднее
  - β_i^Y, β_j^X – главные эффекты: $\beta_*^Y = 0$, $\beta_*^X = 0$
  - β^{YX} – взаимодействия: $\beta_{i*}^{YX} = 0$, $\beta_{*j}^{YX} = 0$
- Гипотеза независимости $H_0 : \beta_{ij}^{YX} = 0, i = 1, \ldots, d_1, j = 1, \ldots, d_2$
  - отсутствие взаимодействий
  - не зависит от выбора весов
  - соответствующая обобщенная линейная модель
$$\log(\lambda_{ij}) = \alpha + \beta_i^Y + \beta_j^X,$$
- β_{ij}^{YX} – логарифмы частных отношений шансов

Сопряженность двух признаков

Проверка гипотезы независимости

$$H_0 : \beta_{ij}^{YX} = 0, i = 1, \dots, d_1, j = 1, \dots, d_2$$

- Критерий типа Вальда
  - H_0 может быть записана с помощью $(d_1 - 1)(d_2 - 1)$ линейно-независимых сравнений $\boldsymbol{\psi} = (\psi_1, \dots, \psi_{(d_1-1)(d_2-1)})'$
  - статистика критерия: $W = \hat{\boldsymbol{\psi}}' \hat{\Gamma}_{\psi}^{-1} \hat{\boldsymbol{\psi}}$
  - значение W не зависит от выбора сравнений
  - предельное распределение: $\chi^2_{(d_1-1)(d_2-1)}$
  - P-значение: pv=$1 - K_{(d_1-1)(d_2-1)}(W)$
- Критерий отношения правдоподобия
  - статистика критерия: $G = 2(LL_S - LL_A)$ (deviance)
    - LL_S максимум логарифма правдоподобия в общей модели
    - LL_A максимум логарифма правдоподобия в аддитивной модели
  - предельное распределение: $\chi^2_{(d_1-1)(d_2-1)}$
  - P-значение: pv=$1 - K_{(d_1-1)(d_2-1)}(G)$
- в отличие от классической модели $W \neq G$ в общем случае

План

Появление дополнительного признака

Парадокс Симпсона

- Статистика осложнений после операции

Больница	Общая		Начальная		Запущенная	
	A	B	A	B	A	B
С осложнениями	66	17	7	8	59	9
Без осложнений	2034	783	593	592	1441	191
Всего	2100	800	600	600	1500	200
Осложнений (%)	3.14%	2.13%	1.17%	1.33%	3.93%	4.50%

  - Вероятности осложнений :
    - общая статистика $p_A = 0.031 > p_B = 0.021$ (приоритет B)
    - начальная стадия $p_A = 0.012 < p_B = 0.013$ (приоритет A)
    - запущенная болезнь $p_A = 0.039 < p_B = 0.045$ (приоритет A)
- Согласно общей статистике лучше лечиться в больнице A
- Углубленный анализ показывает, что в больнице B статистика осложнений благоприятнее, как случае начальной стадии болзни, так и в случае запущенной болезни

Сопряженность трех признаков

Обозначим признаки X, Y, Z:

- $Y \in \{1, \dots, d_1\}$, $X \in \{1, \dots, d_2\}$, $Z \in \{1, \dots, d_3\}$

Модели сопряженности трех признаков

- Все признаки равнозначны:
  - совместное распределение (Y, X, Z)
- Два наблюдаемых и один контролирующий признак:
  - Z – независимая переменная; Y и X – зависимые переменные
  - условное распределение вектора (Y, X) при условии Z
- Один наблюдаемый и два контролирующих признака:
  - X, Z – независимые переменные; Y – зависимая переменная
  - условное распределение вектора Y при условии X, Z

Сопряженность трех признаков

(А) Таблица сопряженности $s \in \{1,\dots,d_3\}$

Y	X: 1	$\dots$	d_2	Всего
1	n_{11s}	$\dots$	n_{1d_2s}	n_{1+s}
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
i	n_{i1s}	$\dots$	n_{id_2s}	n_{i+s}
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
d_1	n_{d_11s}	$\dots$	$n_{d_1d_2s}$	n_{d_1+s}
Всего	n_{+1}	$\dots$	n_{+d_2}	n_{++s}

(Б) Совместное распределение $s = 1,\dots,d_3$

Y	X: 1	$\dots$	d_2	Распр. (Y,Z)
1	p_{11s}	$\dots$	p_{1d_2s}	p_{1+s}
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
i	p_{i1s}	$\dots$	p_{id_2s}	p_{i+s}
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
d_1	p_{d_11s}	$\dots$	$p_{d_1d_2s}$	p_{d_1+s}
Распр. (X,Z)	p_{+1s}	$\dots$	p_{+d_2s}	p_{++s}

(В) Условное распределение $s \in \{1,\dots,d_3\}$

Y	X: 1	$\dots$	d_2	Распр. $Y\|Z$
1	$p_{11\|s}$	$\dots$	$p_{1d_2\|s}$	$p_{1+\|s}$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
i	$p_{i1\|s}$	$\dots$	$p_{id_2\|s}$	$p_{i+\|s}$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
d_1	$p_{d_11\|s}$	$\dots$	$p_{d_1d_2\|s}$	$p_{d_1+\|s}$
Распр. $X\|Z$	$p_{+1\|s}$	$\dots$	$p_{+d_2\|s}$	1

- Массив сопряженности имеет размерность $d_1 \times d_2 \times d_3$
  - сечение массива сопряженности (А) – таблица сопряженности
- Модель совместного распределения (Y, X)
  - элементы массива сопряженности имеют совместное мультиномиальное распределение с параметрами из (Б)
- Модель условного распределения Y при условии X (В)
  - считаем, что $p_{i++} > 0,\ i = 1,\dots,d_1,\ p_{+j+} > 0,\ j = 1,\dots,d_2$ и $p_{++s} > 0,\ s = 1,\dots,d_3$
  - $p_{ij|s} = p_{ijs}/p_{++s},\ i = 1,\dots,d_1,\ j = 1,\dots,d_2,\ s = 1,\dots,d_3$

Сопряженность трех признаков

Гипотезы:

- $H_{CI(Z)}$ – условная независимость (Y, X) при условии Z
  - в модели условного распределения (Y, X) при условии Z

    $H_{CI(Z)} : p_{ij|s} = p_{i+|s}p_{+j|s},\ i = 1, \ldots, d_1, j = 1, \ldots, d_2, s = 1, \ldots, d_3$
  - в модели условного распределения Y при условии X, Z

    $H_{CI(Z)} : p_{i|js} = p_{i|j1},\ i = 1, \ldots, d_1, j = 1, \ldots, d_2, s = 1, \ldots, d_3$
    - $p_{i|js} = p_{ijs}/p_{+js},\ i = 1, \ldots, d_1, j = 1, \ldots, d_2, s = 1, \ldots, d_3$
- $H_{I(Z)}$ – независимость (Y, X) и Z
  - в модели условного распределения (Y, X) при условии Z

    $H_{I(Z)} : p_{ij|s} = p_{ij|1},\ i = 1, \ldots, d_1, j = 1, \ldots, d_2, s = 1, \ldots, d_3$
  - в модели совместного распределения (Y, X, Z)

    $H_{I(Z)} : p_{ijs} = p_{ij+}p_{++s},\ i = 1, \ldots, d_1, j = 1, \ldots, d_2, s = 1, \ldots, d_3$
- H_I – независимость (Y, X, Z) в совокупности
  - в модели совместного распределения (Y, X, Z)

    $H_I : p_{ijs} = p_{i++}p_{+j+}p_{++s},\ i = 1, \ldots, d_1, j = 1, \ldots, d_2, s = 1, \ldots, d_3$
  - $H_I \Leftrightarrow H_{CI(Z)} \cap H_{CI(X)} \cap H_{CI(Y)} \Leftrightarrow H_{I(Z)} \cap H_{I(X)} \cap H_{I(Y)}$

Проверка гипотезы независимости H_I

Критерий χ^2

- статистика критерия: $X^2 = \sum_{i,j,s} \dfrac{(n_{ijs} - \hat{\mu}_{ijs})^2}{\hat{\mu}_{ijs}}$
  - $\hat{\mu}_{ijs} = n\hat{p}_{i++}\hat{p}_{+j+}\hat{p}_{++s} = n_{i++}n_{+j+}n_{++s}/n^2$
- асимптотическое распределение: χ^2_q
  - $q = d_1 d_2 d_3 - d_1 - d_2 - d_3 + 2$
- доверительная область: $X^2 \le x_\alpha,\ x_\alpha : K_q(x_\alpha) = 1 - \alpha$
- P-значение: $pv = 1 - K_q(X^2)$

Критерий отношения правдоподобия

- статистика критерия: $G^2 = 2\sum_{ijs} n_{ijs} \log(n_{ijs}/\hat{\mu}_{ijs})$
- асимптотическое распределение: χ^2_q
- доверительная область: $X^2 \le x_\alpha,\ x_\alpha : K_q(x_\alpha) = 1 - \alpha$
- P-значение: $pv = 1 - K_q(G^2)$

Сопряженность трех признаков

Модель логистической регрессии

- применима для ограниченного круга задач сопряженности (Y, X, Z)
  - $Y \in \{0, 1\}$ – бинарный признак
  - используется условная модель распределения Y при условии (X, Z)
  - Y – наблюдаемый признак; X, Z – контролируемые признаки
- по сути сводится к анализу зависимости Y от X и Z.
- обобщенная линейная модель

$$\operatorname{logit}(\pi_{ij}) = \alpha + \beta_i^X + \beta_j^Z + \beta_{ij}^{XZ}, \; i = 1, \dots, d_1, j = 1, \dots, d_2$$

  - $\pi_{ij} = p_{1|ij} = p_{1ij}/p_{1++}$ – условные вероятности
  - правая часть – двухфакторный дисперсионный анализ
  - α – взвешенное среднее
  - β_i^X; β_j^Z – главные эффекты: $\beta_*^X = 0$, $\beta_*^Z = 0$
  - β_{ij} – взаимодействия: $\beta_{i*} = 0$, $\beta_{*j} = 0$
- гипотеза аддитивности влияния факторов (X, Z) на результат Y

$$H_{A0} : \beta_{ij}^{XZ} = 0 \quad \forall\, i, j$$

- гипотеза независимости Y от (X, Z)

$$H_{I0} : \beta_i^X = 0, \beta_j^Y = 0, \beta_{ij}^{XZ} = 0 \quad \forall\, i, j$$

Сопряженность трех признаков

Лог-линейная модель

- Применима для анализа сопряженнотси признаков (Y, X, Z) общего вида
  - $X \in \{1, \dots, d_1\}$, $Y \in \{1, \dots, d_2\}$, $Z \in \{1, \dots, d_3\}$
- Параметры распределения $\lambda_{ijs} = \lambda p_{ijs}$, $i = 1, \dots, d_1$, $j = 1, \dots, d_2$, $s = 1, \dots, d_3$
- Пуассоновская схема (совместное распределение)
  - n_{ijs} независимы
  - n_{ijs} имеют пуассоновское распределение $\mathrm{Pois}(\lambda_{ijs})$
- Обобщенная линейная модель

$$\log(\lambda_{ij}) = \alpha + \beta_i^Y + \beta_j^X + \beta_j^Z + \beta_{ij}^{YX} + \beta_{is}^{YZ} + \beta_{js}^{XZ} + \beta_{ijs}^{YXZ},$$

  - правая часть – трехфакторный дисперсионный анализ
  - α – взвешенное среднее
  - β_i^Y, β_j^X, β_s^Z – главные эффекты: $\beta_*^Y = 0$, $\beta_*^X = 0$, $\beta_*^Z = 0$
  - β_{ij}^{YX}, β_{is}^{YZ}, β_{js}^{XZ} – взаимодействия: $\beta_{ij*}^{YX} = 0$, $\beta_{i*s}^{YZ} = 0$, $\beta_{*js}^{XZ} = 0$
  - β_{ijs}^{YXZ} – взаимодействия 3-х факторов: $\beta_{ij*}^{YXZ} = \beta_{i*s}^{YXZ} = \beta_{*js}^{YXZ} = 0$

Сопряженность трех признаков

Обобщенная линейная модель

$$\log(\lambda_{ij}) = \alpha + \beta_i^Y + \beta_j^X + \beta_s^Z + \beta_{ij}^{YX} + \beta_{is}^{YZ} + \beta_{js}^{XZ} + \beta_{ijs}^{YXZ},$$

Основные гипотезы

- Однородность зависимостей

$$H_{HA} : \beta_{ijs}^{YXZ} = 0, \quad \forall\ i, j, s$$

  - однородность отношений шансов в случае трех бинарных признаков
  - однородность зависимостей не зависит от выбора контролируемого фактора

- Условная независимость (Y, X) при условии Z

$$H_{CI} : \beta_{ij}^{YX} = 0, \beta_{ijs}^{YXZ} = 0, \quad \forall\ i, j, s$$

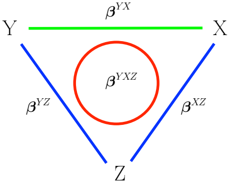
- Независимость (Y, X, Z)

$$H_I : \beta_{ij}^{YX} = 0, \beta_{is}^{YZ} = 0, \beta_{js}^{XZ} = 0, \beta_{ijs}^{YXZ} = 0, \quad \forall\ i, j, s$$

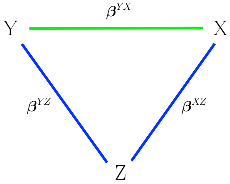
  - аддитивная модель

Сопряженность трех признаков

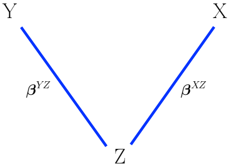
Общие предположения



Однородность зависимостей



Условная независимость



Независимость