

САНИ

Минкин Даниэль

18 января 2025 г.

Содержание

1 Введение	1
1.1 Формула оценки	1
2 Лекция 1	1
2.1 Виды шкал	1
2.1.1 Номинальная шкала	3
2.1.2 Порядковая шкала	3
2.1.3 Интервальная шкала	4
2.1.4 Шкала разностей	4
2.1.5 Шкала отношений	5
2.1.6 Абсолютная шкала	5
2.2 С чем будем работать на САНИ?	6
2.3 Проверка независимости двух дихотомических признаков . .	6
2.4 Независимость в таблицах сопряженности	7
2.4.1 МНП-оценка	7
2.4.2 Случай независимости	8
2.4.3 Асимптотическая проверка независимости	9
2.4.4 Точная проверка независимости	11

1 Введение

1.1 Формула оценки

$$0.1 * \text{Квизы} + 0.4 * \text{Дз} + 0.5 * \text{Экзамен} \quad (1)$$

Курс включает в себя 4 дз

Квизы проводятся на лекциях и в конце семинаров

2 Лекция 1

2.1 Виды шкал

Выделяют следующие виды шкал:

- Номинальная шкала
- Порядковая шкала
- Интервальная шкала
- Шкала разностей
- Шкала отношений
- Абсолютная шкала

Шкала измерения — гомоморфное отображение множества элементов системы с отношениями в множество с заданными логическими отношениями

Вспомним, что такое гомоморфное отношение

Гомоморфное отношение (отображение) — отображение сохраняющее свойства заданные на первом множестве, после его отображения в новое.

Заскочим вперед ради примера: допустим у нас есть результаты опросов с двумя вариантами ответов — “хорошо” и “плохо”. На этом множестве может быть задана операция сравнения, а именно “хорошо” $>$ “плохо”, тогда при переходе к шкале мы должны перейти к порядковой шкале, такой, что $f(\text{“хорошо”}) > f(\text{“плохо”})$, например обозначать “хорошо” как 1, а “плохо” как 0

Рассмотрим подробнее свойства шкал, которые мы будем рассматривать далее:

- **Тождество** — на множестве элементов шкалы задана операция равенства
- **Порядок** — на множестве элементов шкалы задана операция
- **Нулевая точка** — точка, с которой начинается отсчет в шкале
- **Единицы измерения** — no comments
- **Операция сложения, вычитания** — не нуждается в комментариях))
- **Операция деления, умножения** — так же не нуждается
- **Допустимое преобразование** — какое преобразование мы можем выполнить с элементами шкалы, оставаясь в ней же
- **Мода** — задана ли на множестве элементов шкалы операция поиска моды

- **Медиана** — задана ли на множестве элементов шкалы операция поиска медианы
- **Ср. арифм и хар-ки рассеяния** — задана ли на множестве элементов шкалы операции поиска среднего и характеристик разброса

2.1.1 Номинальная шкала

- **Тождество** — Да
- **Порядок** — Нет
- **Нулевая точка** — Нет
- **Единицы измерения** — Нет
- **Операция сложения, вычитания** — Нет
- **Операция деления, умножения** — Нет
- **Допустимое преобразование** — Взаимно-однозначное
- **Мода** — Да
- **Медиана** — Нет
- **Ср. арифм и хар-ки рассеяния** — Нет

Примеры: пол, номера паспортов, ИНН

2.1.2 Порядковая шкала

- **Тождество** — Да
- **Порядок** — Да
- **Нулевая точка** — Может существовать
- **Единицы измерения** — Нет
- **Операция сложения, вычитания** — Нет
- **Операция деления, умножения** — Нет
- **Допустимое преобразование** — Строго монотонное
- **Мода** — Да
- **Медиана** — Да
- **Ср. арифм и хар-ки рассеяния** — Нет

Примеры: оценки успеваемости, рейтинг облигаций

2.1.3 Интервальная шкала

- Тожество — Да
- Порядок — Да
- Нулевая точка — Опционально, но да
- Единицы измерения — Опционально, но да
- Операция сложения, вычитания — Да
- Операция деления, умножения — Нет
- Допустимое преобразование — Линейное
- Мода — Да
- Медиана — Да
- Ср. арифм и хар-ки рассеяния — Да

Примеры: Шкала Цельсия

2.1.4 Шкала разностей

- Тожество — Да
- Порядок — Да
- Нулевая точка — Опционально, да
- Единицы измерения — Однозначно определены
- Операция сложения, вычитания — Да
- Операция деления, умножения — Нет
- Допустимое преобразование — Сдвиг
- Мода — Да
- Медиана — Да
- Ср. арифм и хар-ки рассеяния — Да

Примеры: Время

2.1.5 Шкала отношений

- Тожество — Да
- Порядок — Да
- Нулевая точка — Однозначно определена
- Единицы измерения — Опционально, но да
- Операция сложения, вычитания — Да
- Операция деления, умножения — Да
- Допустимое преобразование — Подобие $f(x) = a * x$
- Мода — Да
- Медиана — Да
- Ср. арифм и хар-ки рассеяния — Да

Примеры: шкала Кельвина, масса тела и длина

2.1.6 Абсолютная шкала

- Тожество — Да
- Порядок — Да
- Нулевая точка — Однозначно определена
- Единицы измерения — Однозначно определены
- Операция сложения, вычитания — Да
- Операция деления, умножения — Да
- Допустимое преобразование — Тожественное $f(x) = x$
- Мода — Да
- Медиана — Да
- Ср. арифм и хар-ки рассеяния — Да

Примеры: Число предметов, событий

2.2 С чем будем работать на САНИ?

Ключевые факты:

- В основном дискретные генеральные совокупности (конечные случайные величины или векторы)
- Зачастую на вход нам подаются данные представленные в виде частот наблюдений, попавших в некоторые категории (или классы)
- Основная шкала измерения - номинальная

Основная задача САНИ — изучение связей, между различными качественными признаками многомерной генеральной совокупности

2.3 Проверка независимости двух дихотомических признаков

Во-первых обозначим, что это за признаки

Дихотомические переменные — а.к.а бинарные, переменные, которые принимают всего два значения

	Y_1	Y_2	p_X
X_1	p_{11}	p_{12}	$p_{1*} = p_{11} + p_{12}$
X_2	p_{21}	p_{22}	$p_{2*} = p_{21} + p_{22}$
p_Y	$p_{*1} = p_{11} + p_{21}$	$p_{*2} = p_{12} + p_{22}$	$p_{**} = p_{11} + p_{12} + p_{21} + p_{22} = 1$

Таблица 1: Вероятностная таблица сопряженности

где p_{ij} — вероятность, того, что случайно взятый объект совокупности обладает категориями X_i и Y_j

	Y_1	Y_2	p_X
X_1	n_{11}	n_{12}	$n_{1*} = n_{11} + n_{12}$
X_2	n_{21}	n_{22}	$n_{2*} = n_{21} + n_{22}$
n_Y	$n_{*1} = n_{11} + n_{21}$	$n_{*2} = n_{12} + n_{22}$	$n_{**} = n_{11} + n_{12} + n_{21} + n_{22} = n_{\text{сумм.}}$

Таблица 2: Частотная таблица сопряженности

где n_{ij} — число элементов выборки, которое обладает категориями X_i и Y_j

2.4 Независимость в таблицах сопряженности

Сформулируем нулевую гипотезу о независимости

Изначально условие задается так:

$$\begin{cases} p_{11} = p_{1*} * p_{*1} \\ p_{12} = p_{1*} * p_{*2} \\ p_{21} = p_{2*} * p_{*1} \\ p_{22} = p_{2*} * p_{*2} \end{cases}$$

Заметим, что выражения полностью эквивалентны. Если выполняется одно из них, то выполняются все остальные. Оставим только одно выражение

$$\begin{aligned} p_{11} &= p_{1*} \cdot p_{*1} \\ \Rightarrow \\ p_{11} &= (p_{11} + p_{12}) \cdot (p_{11} + p_{21}) \\ \Rightarrow \\ p_{11} &= p_{11}^2 + p_{11} \cdot p_{21} + p_{12} \cdot p_{11} + p_{12} \cdot p_{21} \\ \Rightarrow \\ p_{11}^2 + p_{11} \cdot p_{21} + p_{12} \cdot p_{11} + p_{12} \cdot p_{21} - p_{11} &= 0 \\ \Rightarrow \\ p_{11} \cdot (p_{11} + p_{21} + p_{12}) + p_{12} \cdot p_{21} - p_{11} &= 0 \\ \Rightarrow \\ p_{11} \cdot (1 - p_{22}) + p_{12} \cdot p_{21} - p_{11} &= 0 \\ \Rightarrow \\ p_{12} \cdot p_{21} - p_{11} \cdot p_{22} &= 0 \end{aligned}$$

Следовательно, критерием независимости является

$$p_{12} \cdot p_{21} = p_{11} \cdot p_{22} \quad (2)$$

2.4.1 МНП-оценка

Зададим функцию наибольшего правдоподобия

$$P(n_{11}, n_{12}, n_{21}, n_{22}) = \frac{(n_{11} + n_{12} + n_{21} + n_{22})!}{n_{11}! \cdot n_{12}! \cdot n_{21}! \cdot n_{22}!} \cdot p_{11}^{n_{11}} \cdot p_{12}^{n_{12}} \cdot p_{21}^{n_{21}} \cdot p_{22}^{n_{22}} \quad (3)$$

где n_{ij} — наблюдаемое знач СВ, а p_{ij} — параметры распределения

Найдем логарифмическую функцию правдоподобия

$$L(n_{11}, n_{12}, n_{21}, n_{22}) = \ln((n_{11} + n_{12} + n_{21} + n_{22})!) - \ln(n_{11}!) - \ln(n_{12}!) - \ln(n_{21}!) - \ln(n_{22}!) + n_{11} \cdot \ln(p_{11}) + n_{12} \cdot \ln(p_{12}) + n_{21} \cdot \ln(p_{21}) + n_{22} \cdot \ln(p_{22}) \quad (4)$$

Вычтем const для более удобной работы

$$L(n_{11}, n_{12}, n_{21}, n_{22}) = n_{11} \cdot \ln(p_{11}) + n_{12} \cdot \ln(p_{12}) + n_{21} \cdot \ln(p_{21}) + n_{22} \cdot \ln(p_{22}) \quad (5)$$

Найдем функцию Лагранжа

$$\mathcal{L} = L + \lambda \cdot (1 - \sum_{i=1}^2 \sum_{j=1}^2 p_{ij}) \quad (6)$$

Следовательно:

$$\begin{cases} \frac{n_{11}}{p_{11}} - \lambda = 0 \\ \frac{n_{12}}{p_{12}} - \lambda = 0 \\ \frac{n_{21}}{p_{21}} - \lambda = 0 \\ \frac{n_{22}}{p_{22}} - \lambda = 0 \\ \sum_{i=1}^2 \sum_{j=1}^2 p_{ij} = 1 \end{cases}$$

$$\begin{cases} \frac{n_{11}}{p_{11}} = \frac{n_{12}}{p_{12}} = \frac{n_{21}}{p_{21}} = \frac{n_{22}}{p_{22}} \\ \sum_{i=1}^2 \sum_{j=1}^2 p_{ij} = 1 \end{cases}$$

Следовательно:

$$\widehat{p_{ij}} = \frac{n_{ij}}{n_{11} + n_{12} + n_{21} + n_{22}} \quad (7)$$

Несколько фактов:

- Безусловные оценки частот, по МНП и ММ совпадают и равны отношению частот n_{ij} к общему объему выборки
- Безусловные оценки частот, являются несмещенными
- При больших объемах выборки и отсутствии малых частот, в соответствии с ЦПТ совместное распределение n_{ij} стремится к многомерному нормальному распределению

2.4.2 Случай независимости

Рассмотрим ситуацию при которой принимается гипотеза о независимости. Тогда:

$$\widehat{p_{ij}} = \frac{n_{i*} \cdot n_{*j}}{n_{**}^2} \quad (8)$$

2.4.3 Асимптотическая проверка независимости

Пирсон

Для асимптотического тестирования используется критерий согласия χ^2 . Используемая статистика:

$$\chi^2 = \sum_{i=1}^2 \sum_{j=1}^2 \frac{(n_{ij} - n_{ij}^*)^2}{n_{ij}^*} \quad (9)$$

где $n_{ij}^* = \frac{n_{i*} \cdot n_{*j}}{n_{**}}$

Крит. область: $\chi_{\text{набл}}^2 > \chi_{\text{кр}}^2$ где $\chi_{\text{кр}}^2 = \min\{x : P(\chi^2(1) > x) < a\}$

Почему берется 1 степень свободы?

Во-первых, начнем с того, что у нас 4 слагаемых и 4 параметра: $p_{11}, p_{12}, p_{21}, p_{22}$. Мы можем сразу заметить, что зная 3 параметра, мы точно находим 4, следовательно кол-во “неизвестных” параметров ≤ 3 . Нужно заметить, что мы проверяем гипотезу принадлежности распределения вектора (X, Y) к распределению, где X и Y независимы, а у такого распределения достаточно знать всего две p_{ij} для того, чтобы вывести все остальные p -шки. \Rightarrow у нас получается $4 - 2 - 1 = 1$ степеней свободы. Ч.Т.Д

При попадении в критическую область нулевая гипотеза отвергается с вероятностью ошибки a

Рассмотрим вывод формулы критерия согласия, для этого возьмем отклонение суммы частот за вычетом 1, домноженное на n_{**}

Нам нужно привести формулу, к такому виду:

$$n_{**} \cdot \left(\sum_{i=1}^2 \sum_{j=1}^2 \frac{n_{ij}^2}{n_{i*} \cdot n_{*j}} - 1 \right) \quad (10)$$

Приступим:

$$\begin{aligned}
& \sum_{i=1}^2 \sum_{j=1}^2 \frac{(n_{ij} - \frac{n_{i*} \cdot n_{*j}}{n_{**}})^2 \cdot n_{**}}{n_{i*} \cdot n_{*j}} \\
&= \sum_{i=1}^2 \sum_{j=1}^2 \frac{(\frac{n_{ij} \cdot n_{**} - n_{i*} \cdot n_{*j}}{n_{**}})^2 \cdot n_{**}}{n_{i*} \cdot n_{*j}} \\
&= \sum_{i=1}^2 \sum_{j=1}^2 \frac{(n_{ij} \cdot n_{**} - n_{i*} \cdot n_{*j})^2}{n_{i*} \cdot n_{*j} \cdot n_{**}} \\
&= \sum_{i=1}^2 \sum_{j=1}^2 \frac{(n_{ij} \cdot n_{**})^2}{n_{i*} \cdot n_{*j} \cdot n_{**}} \\
&- 2 \cdot \sum_{i=1}^2 \sum_{j=1}^2 \frac{(n_{i*} \cdot n_{*j}) \cdot (n_{ij} \cdot n_{**})}{n_{i*} \cdot n_{*j} \cdot n_{**}} \\
&+ \sum_{i=1}^2 \sum_{j=1}^2 \frac{(n_{i*} \cdot n_{*j})^2}{n_{i*} \cdot n_{*j} \cdot n_{**}} \\
&= \sum_{i=1}^2 \sum_{j=1}^2 \frac{n_{ij}^2 \cdot n_{**}}{n_{i*} \cdot n_{*j}} \\
&- 2 \cdot n_{**} \\
&+ \sum_{i=1}^2 \sum_{j=1}^2 \frac{n_{i*} \cdot n_{*j}}{n_{**}}
\end{aligned}$$

Рассмотрим, последнее слагаемое:

$$\begin{aligned}
& \frac{n_{1*} \cdot n_{*1} + n_{1*} \cdot n_{*2} + n_{2*} \cdot n_{*1} + n_{2*} \cdot n_{*2}}{n_{**}} \\
&= \frac{n_{1*} \cdot (n_{*1} + n_{*2}) + n_{2*} \cdot (n_{*1} + n_{*2})}{n_{**}} \\
&= \frac{n_{**}^2}{n_{**}} \\
&= n_{**}
\end{aligned}$$

Следовательно, наше выражение имеет вид:

$$\begin{aligned}
& \sum_{i=1}^2 \sum_{j=1}^2 \frac{n_{ij}^2 \cdot n_{**}}{n_{i*} \cdot n_{*j}} \\
& - 2 \cdot n_{**} \\
& + n_{**} \\
& = \sum_{i=1}^2 \sum_{j=1}^2 \frac{n_{ij}^2 \cdot n_{**}}{n_{i*} \cdot n_{*j}} - n_{**} \\
& = n_{**} \cdot \left(\sum_{i=1}^2 \sum_{j=1}^2 \frac{n_{ij}^2}{n_{i*} \cdot n_{*j}} - 1 \right)
\end{aligned}$$

Ч.Т.Д

Критерий отношения правдоподобия

Рассмотрим статистику критерия:

$$\chi_{\text{инф}}^2 = 2 \cdot \sum_{i=1}^2 \sum_{j=1}^2 n_{ij} \cdot \ln\left(\frac{n_{ij}}{n_{ij}^*}\right) \quad (11)$$

где $n_{ij}^* = \frac{n_{i*} \cdot n_{*j}}{n_{**}}$ получается, что n_{ij}^* — теоретическая частота

2.4.4 Точная проверка независимости

Гипотеза независимости $H_0 : p_{11} \cdot p_{22} = p_{12} \cdot p_{21}$

Крит. обл.