САНИ

Минкин Даниэль

17 января 2025 г.

Содержание

1 Введение

1.1 Формула оценки

$$0.1 * \text{Квизы} + 0.4 * \text{Дз} + 0.5 * \text{Экзамен} \tag{1}$$

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(“хорошо“) > f(“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

- **Тождество** — 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

- **Тождество** — 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

- **Тождество** — 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

- **Тождество** — Да
- **Порядок** — Да
- **Нулевая точка** — Однозначно определена
- **Единицы измерения** — Однозначно определены
- **Операция сложения, вычитания** — Да
- **Операция деления, умножения** — Да
- **Допустимое преобразование** — Тождественное $f(x) = x$
- **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** — a.k.a бинарные, переменные, которые принимают всего два значения

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

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

	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

	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

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

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

Следовательно, критерием независимости является

$$p_{12} \cdot p_{21} = p_{11} \cdot p_{22} \tag{2}$$

2.4.1 МНП-оценка

Зададим функцию наибольшего правдоподобия

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

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

Найдем логарифмическую функцию правдоподобия

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

Вычтем const для более удобной работы

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

Найдем функцию Лагрнажа

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

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:

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

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:

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

2.4.3 Асимптотическая проверка независимости

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

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

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

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

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

2.4.4 Точная проверка независимости

Гипотеза независимости $H0 : p_{11} \cdot p_{22} = p_{12} \cdot p_{21}$

Пусть маргинальные частоты зафиксированы: i.e