Продвинутые количественные методы Анализа данных

Парамей Мария
CERGE-EI, PhD Candidate

@mparamei

Анализ данных

Процесс изучения, очищения, трансформации и моделирования данных с целью получения полезной информации, формирования выводов и мотивирования принятия решений.

- Сбор
- Структурирование
- Очищение
- Эксплораторный анализ
- Моделирование

Анализ данных

Анализ данных – процесс изучения, очищения, трансформации и моделирования данных с целью получения полезной информации, формирования выводов и мотивирования принятия решений.

- Сбор
- Структурирование
- Очищение
- **Эксплораторный анализ**
- Моделирование

Эксплораторный («разведочный») анализ

- Центральность
- Вариация
- Взаимосвязи
- Визуализация

Обобщающие статистические показатели

Средняя величина - обобщающая характеристика множества индивидуальных значений некоторого количественного признака.

Средняя величина служит мерой признака на единицу совокупности, отражает то общее, что присуще всем единицам исследуемой совокупности.

- средняя арифметическая
- средняя гармоническая
- средняя геометрическая
- средняя квадратическая

Средняя арифметическая

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- Средняя арифметическая постоянной величины равна этой постоянной ($\bar{x} = x$, если $x = const$)
- Сумма отклонений индивидуальных значений признака от средней арифметической равна нулю ($\sum_{i=1}^{n}(x_i - \bar{x})f_i = 0$)
- Сумма квадратов отклонений индивидуальных значений признака от средней арифметической меньше, чем от любого другого числа ($\sum_{i=1}^{n}(x_i - \bar{x})^2 < \sum_{i=1}^{n}(x_i - C)^2$)

Средние величины

$$\bar{x} = \frac{n}{\sum_{i=1}^{n} 1/x_i}$$ средняя гармоническая

$$\bar{x} = \sqrt[n]{\prod_{i=1}^{n} x_i}$$ средняя геометрическая

$$\bar{x} = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n}}$$ средняя квадратическая

Средние величины

Неравенства о средних (Коши):

$$\bar{x}_{\text{гарм}} \leq \bar{x}_{\text{геом}} \leq \bar{x}_{\text{арифм}} \leq \bar{x}_{\text{квадр}}$$

Структурные средние величины

- Медиана – значение, которое находится в середине ранжированного ряда

  Ранг медианы: $r_{me} = \frac{n+1}{2}$

- Мода - наиболее распространенная категория

  Пример: 55,4,8,27,8,30,20,89,72

Структурные средние величины

Показатель	Количественные	Порядковые	Номинальные
Средняя	+	–	–
Медиана	+	+	–
Мода	+	+	+

Другие статистические показатели

- Перцентиль (процентиль) - показатель, разбивающий ранжированную совокупность на 100 равных частей.
- Дециль - показатель, разбивающий ранжированную совокупность на 10 равных частей.
- Квартиль - показатель, разбивающий ранжированную совокупность на 4 равные части.

Показатели вариации

Вариация признака (изменчивость) - степень различий между отдельными значениями признака, неоднородность признака.

- Размах – интервал, занимаемый значениями данных

$$R = r_{max} - r_{min}$$

- Среднее линейное отклонение – число, описывающее, насколько значения данных обычно отличаются от среднего.

$$\bar{d} = \frac{\sum |x_i - \bar{x}|}{n}$$

Показатели вариации

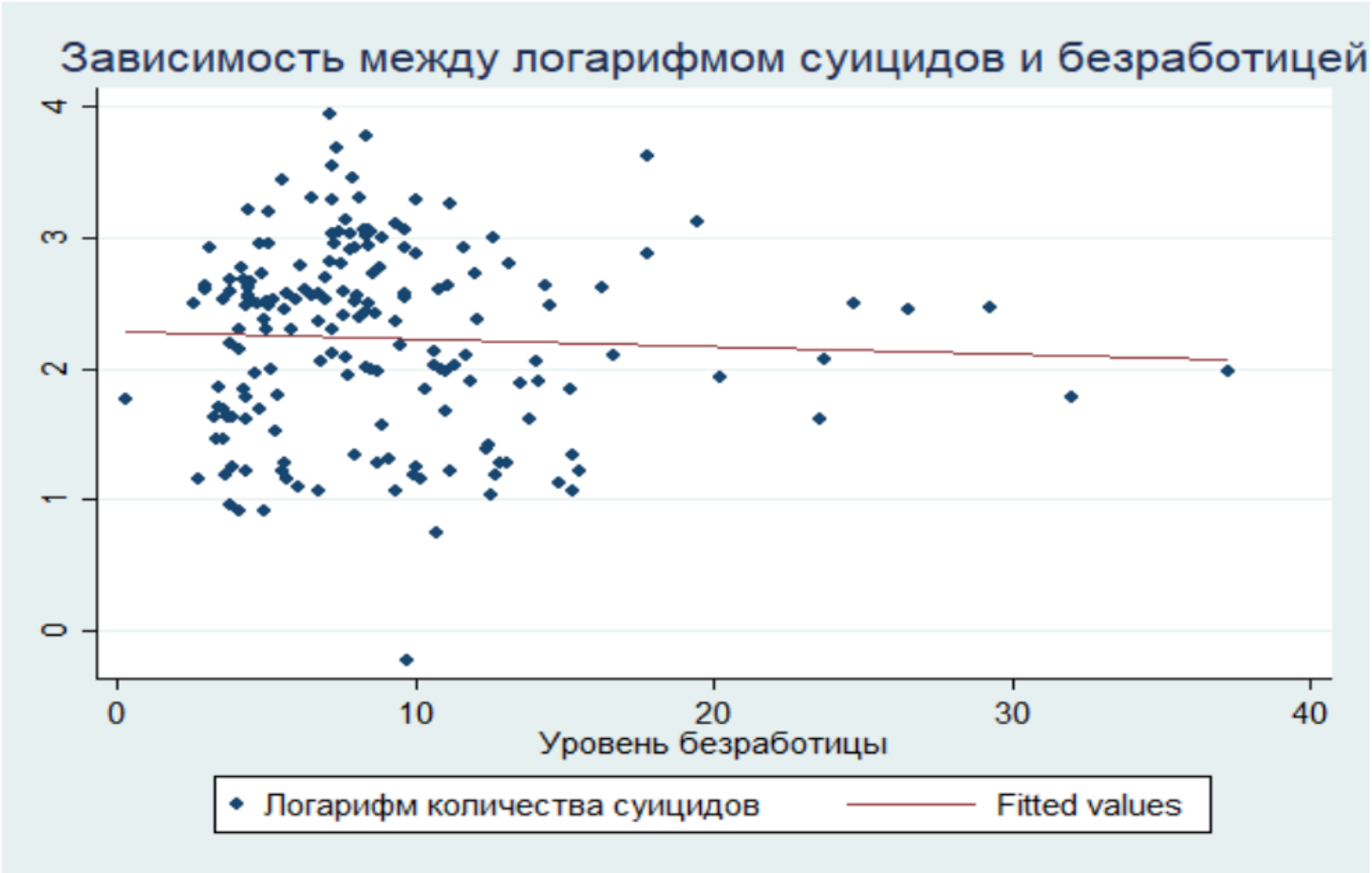
- Среднее квадратическое отклонение – число, описывающее, насколько значения данных обычно отличаются от среднего.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}$$

- Дисперсия – мера разброса данных от среднего значения.

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$$

Взаимосвязь количественных признаков



Взаимосвязь количественных признаков

Линейный (парный) коэффициент корреляции Пирсона

$$r_{xy} = \frac{\sum(x_i - \bar{x}) \times (y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \times (y_i - \bar{y})^2}}$$

Значение r	Характер связи
От 0 до $\lvert\pm 0,3\rvert$	Практически отсутствует
От $\lvert\pm 0,3\rvert$ до $\lvert\pm 0,5\rvert$	Слабая
От $\lvert\pm 0,5\rvert$ до $\lvert\pm 0,7\rvert$	Умеренная
От $\lvert\pm 0,7\rvert$ до $\lvert\pm 1\rvert$	Сильная

Взаимосвязь качественных признаков

a	b	a+b
c	d	c+d
a+c	b+d	a+b+c+d

$$K_{\text{ассоц}} = \frac{ad - bc}{ad + bc}$$

$$K_{\text{конт}} = \frac{ad - bc}{\sqrt{(a+b)(a+c)(c+d)(b+c)}}$$

$$K_{\text{ассоц}} \geq K_{\text{конт}}$$