Статистика – 2.3 Анализ концентрации

Д-р Вольфганг Бёме

Ноябрь 2010

2. Одномерный анализ распределений – Содержание

- 2.1 Распределение частот
- 2.2 Числовые характеристики распределения
- **2.3 Анализ концентрации**

Цель / Содержание

Ранее: меры диспаритета: отклонение распределения от равномерного

Сейчас: исследование связи между общей суммой всех значений признака и отдельными наблюдениями статистической совокупности

Рассматриваются **экстенсивные признаки**, т.е. наблюдения неотрицательны ($x_i \geq 0$) и образование общей суммы значений признака содержательно правомерно.

Содержание:

- 2.3.1 **абсолютная статистическая концентрация**
- 2.3.2 относительная статистическая концентрация

Коэффициент концентрации

Определение: Пусть x_i $(i = 1, 2, ..., n)$ – наблюдения экстенсивного признака, упорядоченные по убыванию $(x_i > x_{i+1})$. Величины C_r называются **абсолютными коэффициентами концентрации**:

$$C_r = \sum_{i=1}^{r} a_i \text{ где}$$

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

Замечание: Получается последовательность абсолютных коэффициентов концентрации C_r (кривая концентрации).
Сравнение с прямой равномерного распределения $(C_r = \frac{r}{n})$.

Интерпретация:

Сильное отклонение от прямой равномерного распределения $\Longrightarrow$ высокая концентрация.

C_r – кумулятивная доля значений признака, упорядоченных по убыванию.



Пример

Объём производства 5 предприятий в млн. евро

Объект	x_i
A	20
B	30
C	15
D	70
E	40

Вопросы:

Какова доля наиболее производительных предприятий?

Как можно интерпретировать коэффициент C_2?

Индекс Херфиндаля

Определение: Пусть x_i $(i = 1, 2, ..., n)$ – наблюдения экстенсивного признака. Величина H называется **индексом Херфиндаля**:

$$H = \sum_{i=1}^{n} a_i^2 = \sum_{i=1}^{n} \left(\frac{x_i}{n \cdot \overline{x}} \right)^2 \text{ где}$$

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

Замечание: $\frac{1}{n} \leq H \leq 1$

Минимальная концентрация: сбыт равномерно распределён между всеми производителями. $H = \frac{1}{n}$

Максимальная концентрация: весь сбыт приходится на одного производителя. $(H = 1)$



Пример

Объём производства 5 объектов в млн. евро

Объект	x_i
A	20
B	30
C	15
D	70
E	40

Вопрос:

Что показывает индекс Херфиндаля?

Пример – интерпретация

$$H = \sum_{i=1}^{n} a_i^2 = 0{,}262040816$$

Интерпретация: $H > 0{,}1 \Rightarrow$ высокая концентрация

Вопрос о слиянии: как изменится степень концентрации при слиянии двух из пяти предприятий?

$$H = \sum_{i=1}^{n} a_i^2 = 0{,}444897959$$

Степень концентрации возрастает.



Содержание

- 2.3.1 абсолютная статистическая концентрация
- 2.3.2 **относительная статистическая концентрация**

Кривая Лоренца

Определение: Пусть x_i $(i = 1, 2, ..., n)$ – наблюдения экстенсивного признака, упорядоченные по возрастанию $(x_i < x_{i+1})$. Кумулятивная доля носителей признака A_i сопоставляется с долями прямой равномерного распределения F_i:

$$A_i = \frac{\sum_{r=1}^{i} x_r}{\sum_{i=1}^{n} x_i}$$

$$F_i = \frac{i}{n}$$

Последовательность (F_i, A_i) $(i = 0, 1, 2, ..., n)$ с $F_0 = A_0 = 0$ называется **кривой ЛОРЕНЦА**.
Замечание: сравнение с прямой равномерного распределения $(F_i = \frac{i}{n})$.

Сильное отклонение от прямой равномерного распределения $\Longrightarrow$ высокая концентрация.



Пример

Объём производства 5 объектов в млн. евро

Объект	x_i
A	20
B	30
C	15
D	70
E	40

Вопрос:

Что показывает кривая Лоренца?



Кривая Лоренца – интерпретация

- Построение графика (F – ось абсцисс, A – ось ординат)
- Прямая через точки (0,0) и (1,1) – прямая равномерного распределения
- Кривая Лоренца: фактическое распределение
- Площадь между кривой Лоренца и прямой равномерного распределения: площадь концентрации
- Чем больше площадь концентрации, тем выше относительная концентрация
- Равномерная концентрация: площадь концентрации $\rightarrow 0$
- Высокая концентрация: площадь концентрации $\rightarrow 0{,}5$



Кривая Лоренца для несовпадающих признаков

Определение: Пусть ξ_i $(i = 1, 2, ..., n)$ – несовпадающие наблюдения экстенсивного признака, упорядоченные по возрастанию $(\xi_i < \xi_{i+1})$, и p_i – соответствующие относительные частоты. Кумулятивная доля носителей признака A_i сопоставляется с долями прямой равномерного распределения F_i:

$$A_i = \frac{\sum_{r=1}^{i} \xi_r p_r}{\sum_{i=1}^{n} \xi_i p_i}$$

$$F_i = \sum_{r=1}^{i} p_r$$

Последовательность (F_i, A_i) $(i = 0, 1, 2, ..., n)$ с $F_0 = A_0 = 0$ называется **кривой ЛОРЕНЦА**.

Замечание: сравнение с прямой равномерного распределения.

Сильное отклонение от прямой равномерного распределения $\Longrightarrow$ высокая концентрация.



Кривая Лоренца для классифицированных / обьединённых признаков

Определение: Пусть x_i^* $(i = 1, 2, \ldots, n)$ или $\overline{x_i}$ $(i = 1, 2, \ldots, n)$ – середины классов или объединённые средние экстенсивного признака, упорядоченные по возрастанию, и p_i – соответствующие относительные частоты. Кумулятивная доля носителей признака A_i сопоставляется с долями прямой равномерного распределения F_i:

$$A_i = \frac{\sum_{r=\mathbf{1}}^{i} x_r^* p_r}{\sum_{i=\mathbf{1}}^{n} x_i^* p_i}$$

$$A_i = \frac{\sum_{r=\mathbf{1}}^{i} \overline{x_r} p_r}{\sum_{i=\mathbf{1}}^{n} \overline{x_i} p_i}$$

$$F_i = \sum_{r=\mathbf{1}}^{i} p_r$$

Последовательность (F_i, A_i) $(i = 0, 1, 2, \ldots, n)$ с $F_{\mathbf{0}} = A_{\mathbf{0}} = 0$ называется **кривой ЛОРЕНЦА**.
Замечание: сравнение площади между прямой равномерного распределения и кривой Лоренца.

Сильное отклонение от прямой равномерного распределения $\Longrightarrow$ высокая концентрация.



Коэффициент ДЖИНИ

Определение: Пусть x_i $(i = 1, 2, ..., n)$ – наблюдения экстенсивного признака, упорядоченные по возрастанию $(x_i < x_{i+1})$. Из кумулятивных долей носителей признака A_i вычисляется **коэффициент ДЖИНИ**:

$$G = 1 - \frac{1}{n} \sum_{i=1}^{n} (A_i + A_{i-1}) \text{ при } A_0 = 0$$

Замечание:

- G близко к 0 $\Longrightarrow$ слабая концентрация
- G близко к 1 $\Longrightarrow$ высокая концентрация

Коэффициент ДЖИНИ – несовпадающие признаки

Определение: Пусть ξ_i $(i = 1, 2, ..., n)$ – несовпадающие наблюдения экстенсивного признака, упорядоченные по возрастанию $(\xi_i < \xi_{i+1})$, и p_i – соответствующие относительные частоты. Из кумулятивных долей носителей признака A_i вычисляется **коэффициент ДЖИНИ**:

$$G = 1 - \sum_{i=1}^{n}(A_i + A_{i-1})p_i \text{ при } A_0 = 0$$

Замечание:

- G близко к 0 $\Longrightarrow$ слабая концентрация
- G близко к 1 $\Longrightarrow$ высокая концентрация
- При объединённом или классифицированном наборе данных вместо несовпадающих значений признака ξ_i используются средние значений классов $\overline{x_i}$ или середины классов x_i^*.

Пример – продолжение

Объём производства 5 объектов в млн. евро

Объект	x_i
A	20
B	30
C	15
D	70
E	40

Коэффициент ДЖИНИ – интерпретация

Коэффициент ДЖИНИ выводится из кривой Лоренца:

$$G = \frac{\text{Площадь между кривой Лоренца и диагональю между (0,0) и (1,1)}}{\text{Площадь ниже диагонали между (0,0) и (1,1)}}$$

$$= \frac{\text{Площадь между кривой Лоренца и диагональю между (0,0) и (1,1)}}{0{,}5}$$

Пример:

Объект	x_i
A	10
B	10
C	30
D	50

Графически коэффициент Джини представляет собой отношение площади между линией равномерного распределения и кривой Лоренца к площади ниже линии равномерного распределения.

Коэффициент ДЖИНИ – интерпретация

