

Статистика – одномерная статистика
2.2 Числовые характеристики распределения

Д-р Вольфганг Бёме

Октябрь 2010

2. Одномерный анализ распределений – Содержание

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

Цель / Содержание

Цель:

- Описание наборов данных с помощью небольшого числа статистических характеристик
- Дальнейшее сжатие представления распределения

Группы характеристик

- 2.2.1 **Меры положения**
- 2.2.2 Меры рассеяния
- 2.2.3 Меры формы
- 2.2.4 Линейные преобразования

2.2.1 Меры положения – Содержание

- **2.2.1.1 Введение**
- 2.2.1.2 Мода
- 2.2.1.3 Медиана
- 2.2.1.4 Квантили
- 2.2.1.5 Пример: Историческое моделирование
- 2.2.1.6 Среднее арифметическое

Меры положения – намерение

Исходная точка: **первичный список** для одномерной статистики

$$X_{(n,1)} = X_{(n)} \begin{pmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{n1} \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

где

$$\begin{aligned} x_{i1} = x_i \quad &- \quad \text{значение у } i\text{-го носителя признака} \\ i \quad &- \quad \text{индекс носителя признака } (i = 1, 2, ..., n) \end{aligned}$$

Цель:

Описание первичного списка с помощью небольшого числа статистических характеристик, которые характеризуют положение значений признака X_i на оси значений признака (кратко X-ось).

Меры положения – принципиальный подход

Требования к значению меры положения:

- возможно более центральное положение
- малое расстояние до наблюдаемых значений признака x_i

$\Rightarrow$
центральное понятие: **расстояние** или **функция расстояния**
Обозначение:

$$d(x_i, a) \quad - \quad \text{функция расстояния}$$

$$a \quad - \quad \text{мера положения}$$
$$x_i \quad - \quad \text{значение признака} \, (i = 1, 2, ..., n)$$

Требование: $F(d(x_i, a)) \rightarrow min$

Функции расстояния

Определение: Отображение $d : X \times X \to \mathbb{R}$ называется **метрикой**, если для произвольных элементов x, y и z из X выполнены следующие аксиомы:

$$\begin{aligned} d(x,x) &= 0 \\ d(x,y) &= 0 \Rightarrow x = y \\ d(x,y) &= d(y,x) \\ d(x,y) &\leq d(x,z) + d(z,y) \end{aligned}$$

Типичные функции расстояния (метрики):

$$\begin{aligned} d(x,y) &= |x-y| \; x,y \in \mathbb{R} \\ d(x,y) &= \sqrt{\sum_{i=1}^{n}(x_i - a)^2}, \;\; x,a \in \mathbb{R} \times \mathbb{R} \times \ldots \times \mathbb{R} \end{aligned}$$

Функции расстояния в одномерной статистике

Типичные функции расстояния (метрики) в одномерной статистике:

$$
\begin{aligned}
d(x,y) &= |x-y| \; x,y \in \mathbb{R} \\
d(x,y) &= (x-a)^2, \;\; x,y \in \mathbb{R} \\
d(x,y) &= \sqrt{(x-a)^2}, \;\; x,y \in \mathbb{R} \\
d(x,a) &= \begin{cases} 0 & \text{если } x = a \\ 1 & \text{если } x \neq a \end{cases}
\end{aligned}
$$

Функции расстояния – пример

Пример: Пример из лекции 1
Лист: Abstandsfunktion_base

Задание: определить расстояния для метрических переменных



2.2.1 Меры положения – Содержание

- 2.2.1.1 Введение
- **2.2.1.2 Мода**
- 2.2.1.3 Медиана
- 2.2.1.4 Квантили
- 2.2.1.5 Пример: Историческое моделирование
- 2.2.1.6 Среднее арифметическое

Мода – определение

Определение: Каждое несовпадающее значение признака ξ_{j*}, абсолютная частота которого

$$n_{j*} = max\{n_1, n_2, ..., n_n\}$$

или относительная частота которого

$$p_{j*} = max\{p_1, p_2, ..., p_n\}$$

называется **модой** (модальным значением).

Обозначение: X_M

Примечание: если специально не оговорено, в дальнейшем несовпадающие значения признака ξ_i кратко называются значениями признака и обозначаются x_i.

Мода – замечания

1. Применима для любых шкал
2. Мода – это значение признака в наборе данных произвольно масштабированных признаков, которое наблюдается чаще всего,
3. В столбчатой диаграмме мода соответствует значению с наибольшим столбцом,
4. Мода не обязательно однозначна,
5. Используемая функция расстояния:

$$d(x_i, a) = \begin{cases} 0 & \text{если } x_i = a \\ 1 & \text{если } x_i \neq a \end{cases}$$

Пример из лекции 1

	Несовп. знач. признака	Абс. частоты	Отн. частоты
j	ξ_j	n_j	p_j
1	1	1	0.02
2	2	4	0.08
3	3	11	0.22
4	4	23	0.46
5	5	8	0.16
6	6	2	0.04
7	7	1	0.02
Сумма		50	1

$$n_{j^*} = max\{n_1, ..., n_n\} = 23 \Rightarrow X_M = 4$$

или относительная частота

$$p_{j^*} = max\{p_1, ..., p_n\} = 0{,}46 \Rightarrow X_M = 4$$

Мода с группировкой по интервалам – определение

Исходная точка: интервал $I_l = (c_{l-1}, c_l)$

Проблема: наиболее часто встречающийся класс легко найти, но какое значение использовать в качестве моды?

2 приближённых метода:

1. Середина класса интервала $I_l \Rightarrow$

$$X_M \approx \frac{1}{2}(c_{l-1} + c_l)$$

2. Мода при допущении равномерного распределения в интервалах I_{l-1}, I_l и $I_{l+1} \Rightarrow$

$$X_M \approx c_{l-1} + \frac{p_l^D - p_{l-1}^D}{(p_l^D - p_{l-1}^D) + (p_l^D - p_{l+1}^D)} \cdot \Delta_l$$

Пример из лекции 2

Häufigkeitsverteilung, tabellarisch

j	x_j^u	x_j^o	n_j	p_j
1	0	2	58	0,58
2	2	4	22	0,22
3	4	6	16	0,16
4	6	10	3	0,03
m=5	10	14	1	0,01
Σ			n=100	1,00



Klassenunter- bzw. -oberarenze

Δ_j	p_j^D
2	0,2900
2	0,1100
2	0,0800
4	0,0075
4	0,0025

Наиболее часто встречающийся класс: $[0; 2)$

$$X_M \approx \frac{1}{2}(c_{l-\mathbf{1}} + c_l) = 1$$

$$X_M \approx c_{l-\mathbf{1}} + \frac{p_l^D - p_{l-\mathbf{1}}^D}{(p_l^D - p_{l-\mathbf{1}}^D) + (p_l^D - p_{l+\mathbf{1}}^D)} \cdot \Delta_l$$

$$X_M \approx 0 + \frac{0{,}29 - 0}{(0{,}29 - 0) + (0{,}29 - 0{,}11)} \cdot 2 = 1{,}234$$



2.2.1 Меры положения – Содержание

- 2.2.1.1 Введение
- 2.2.1.2 Мода
- **2.2.1.3 Медиана**
- 2.2.1.4 Квантили
- 2.2.1.5 Пример: Историческое моделирование
- 2.2.1.6 Среднее арифметическое

Медиана – определение

Пусть задан ранговый ряд набора данных, измеренного хотя бы по порядковой шкале:

$$x_1 \leq x_2 \leq \ ..., \ x_n$$

Определение: **Медиана** $\widetilde{x}$ порядкового набора данных при нечётном числе наблюдений:

$$\widetilde{x} = x_{\frac{n+1}{2}}$$

Интерпретация: не менее 50% наблюдений меньше или равны $\widetilde{x}$ и не менее 50% всех наблюдений больше или равны $\widetilde{x}$.

Медиана – определение

Пусть задан ранговый ряд набора данных, измеренного хотя бы по порядковой шкале:

$$x_1 \leq x_2 \leq \ldots, x_n$$

Определение: **Медиана** $\widetilde{x}$ при чётном числе наблюдений:

$$\begin{aligned} \widetilde{x} &\in \left\{ x_{\frac{n}{2}}, x_{\frac{n}{2}+1} \right\} \\ \text{часто} & \\ &= \frac{1}{2}(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}) \end{aligned}$$

Интерпретация: не менее 50% наблюдений меньше или равны $\widetilde{x}$ и не менее 50% всех наблюдений больше или равны $\widetilde{x}$.

Пример из лекции 1

	Несовп. знач. признака	Абс. частоты	Отн. частоты
j	ξ_j	n_j	p_j
1	1	1	0.02
2	2	4	0.08
3	3	11	0.22
4	4	23	0.46
5	5	8	0.16
6	6	2	0.04
7	7	1	0.02
Сумма		50	1

Число наблюдений равно 50 (чётное число)

$$\begin{aligned} \widetilde{x} &\in (x_{\frac{n}{2}}, x_{\frac{n}{2}+1}) \\ \widetilde{x} &\in (x_{25}, x_{26}) \\ \widetilde{x} &\in (4,4) = 4 \end{aligned}$$

часто

$$\widetilde{x} = \frac{1}{2}(x_{25} + x_{26}) = \frac{1}{2}(4+4) = 4$$

Медиана – обобщение

Пусть задан ранговый ряд порядкового набора данных:

$$x_1 \leq x_2 \leq \ldots, x_n$$

p^* – заданное значение (для медианы $p^* = 0{,}5$)

Интерпретация: не менее $p^* \cdot 100\%$ наблюдений меньше или равны $\widetilde{x}$ и не менее $p^* \cdot 100\%$ всех наблюдений больше или равны $\widetilde{x}$

Разбиение рангового ряда на 2 части:

$$\underbrace{x_1 \leq x_2 \leq \ldots, x_k}_{p^* \cdot 100\%} \quad \underbrace{\leq x_{k+1} \leq x_{k+2} \leq \ldots, x_k}_{(1-p^*) \cdot 100\%}$$

Это обобщение медианы называется **квантилем**.



2.2.1 Меры положения – Содержание

- 2.2.1.1 Введение
- 2.2.1.2 Мода
- 2.2.1.3 Медиана
- **2.2.1.4 Квантили**
- 2.2.1.5 Пример: Историческое моделирование
- 2.2.1.6 Среднее арифметическое

Квантиль – определение

Пусть задан ранговый ряд набора данных, измеренного хотя бы по порядковой шкале:

$$x_1 \leq x_2 \leq \ ..., \leq x_n$$

Определение: Пусть X – хотя бы порядковый, численно зафиксированный и упорядоченный признак статистической совокупности, а p, $0 < p < 1$, – заданное значение относительной кумулятивной частоты (эмпирической функции распределения). Тогда наименьшее значение x, для которого $F(x) \geq p$, называется **квантилем порядка p**. *Обозначение*: x_p *Формально*:

$$x_p = min\,(x \mid F(x) \geq p)$$

Интерпретация: $p \cdot 100\%$ всех наблюдений имеют значение, не превышающее (x_p), или максимальное значение, которое могут принимать $p \cdot 100\%$ всех наблюдений, равно (x_p).

Квантиль – определение из отдельных значений

Пусть задан ранговый ряд набора данных, измеренного хотя бы по порядковой шкале:

$$x_1 \leq x_2 \leq \ldots \leq x_n$$

Вычислить значение $n \cdot p$

- если $g = n \cdot p$ – целое, то $x_p = x_g$
- иначе взвешивание соседних целых значений x_g и x_{g+1}:

$$\begin{aligned} \text{если } g &< n \cdot p \leq g + 1 \\ \text{то } x_p &= (1 - r) \cdot x_g + r \cdot x_{g+1} \\ \text{где } n \cdot p &= g + r \ (0 < r < 1,\ r - \text{дробная часть } n \cdot p) \end{aligned}$$

Квантиль – определение из отдельных значений

Задание: Определить 0,2-квантиль ($p = 0{,}2$), т.е. 20-процентный квантиль для примера из лекции 1

Содержательный вопрос: Какова максимальная заселённость квартиры, достигаемая 20% всех квартир?

	Несовп. знач. признака	Абс. частоты	Отн. частоты
j	ξ_j	n_j	p_j
1	1	1	0.02
2	2	4	0.08
3	3	11	0.22
4	4	23	0.46
5	5	8	0.16
6	6	2	0.04
7	7	1	0.02
Сумма		50	1

Число наблюдений равно 50 ($n = 50$)

Квантиль – определение из отдельных значений

Задание: Определить 0,20-квантиль ($p = 0{,}20$) для примера из лекции 1

Содержательный вопрос: Какова максимальная заселённость квартиры, достигаемая 20% всех квартир?

Определение квантиля:

$$\begin{aligned} g &= n \cdot p = 50 \cdot 0{,}20 = 10 \ \rightarrow \text{целое “номер наблюдения”} \\ x_{20} &= 3 \end{aligned}$$

Интерпретация:

1. Не менее 20% всех квартир заселены не более чем 3 лицами.
2. Максимальная заселённость, достигаемая 20% всех квартир, равна 3.
3. Не более 80% всех квартир заселены не менее чем 3 лицами.

Квантиль – определение из отдельных значений – заселённость квартир

Определение 5-процентного квантиля:

$$\begin{aligned} g + r &= n \cdot p = 50 \cdot 0{,}05 = 2{,}5 \ \rightarrow \text{нецелое} \\ g &= 2 \\ r &= 0{,}5 \\ x_g &= x_2 = 2 \\ x_{g+1} &= x_3 = 2 \\ x_p &= 0{,}5x_2 + 0{,}5x_3 = 2 \end{aligned}$$

Интерпретация:

Максимальная заселённость, достигаемая 5% всех квартир, равна 2.

Квантиль – определение из отдельных значений – телефонные разговоры

Пример из лекции 2 (Телефонные разговоры)
Определение 1,25-процентного квантиля:

$$
\begin{aligned}
g &= n \cdot p = 100 \cdot 0{,}0125 = 1{,}25 \ \rightarrow \text{нецелое} \\
g &= 1 \\
r &= 0{,}25 \text{ дробная часть} \\
x_1 &= 0{,}04 \\
x_2 &= 0{,}09 \\
x_p &= (1-r) \cdot x_g + r \cdot x_{g+1} \\
x_{0{,}0125} &= (1-0{,}25) \cdot x_1 + 0{,}25 \cdot x_2 \\
x_{0{,}0125} &= 0{,}75 \cdot 0{,}04 + 0{,}25 \cdot 0{,}09 = 0{,}0525
\end{aligned}
$$

Интерпретация:

Максимальная продолжительность телефонного разговора, достигаемая 1,25% всех разговоров, составляет 0,0525 мин.

Пример из лекции 1

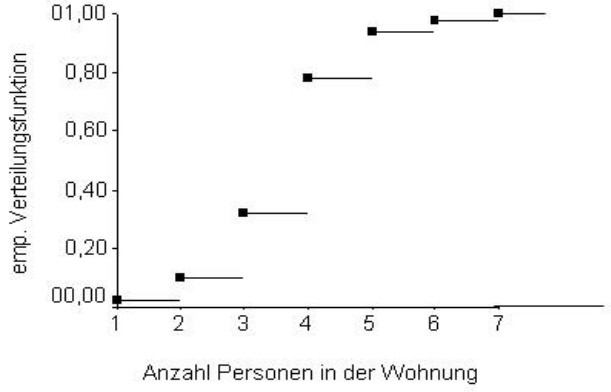
Urliste **aufsteigend sortierte Urliste**

$$\begin{pmatrix} x_1 \\ x_2 \\ . \\ x_i \\ . \\ x_n \end{pmatrix} \Rightarrow \begin{pmatrix} x_{(1)} \\ x_{(2)} \\ . \\ x_{(i)} \\ . \\ x_{(n)} \end{pmatrix} \Downarrow$$

Häufigkeitsverteilung, tabellarisch

j	ξ_j	n_j	p_j	H_j	F_j
1	1	1	0,02	1	0,02
2	2	4	0,08	5	0,10
3	3	11	0,22	16	0,32
4	4	23	0,46	39	0,78
5	5	8	0,16	47	0,94
6	6	2	0,04	49	0,98
m=7	7	1	0,02	50	1,00
Σ		n= 50	1,00		

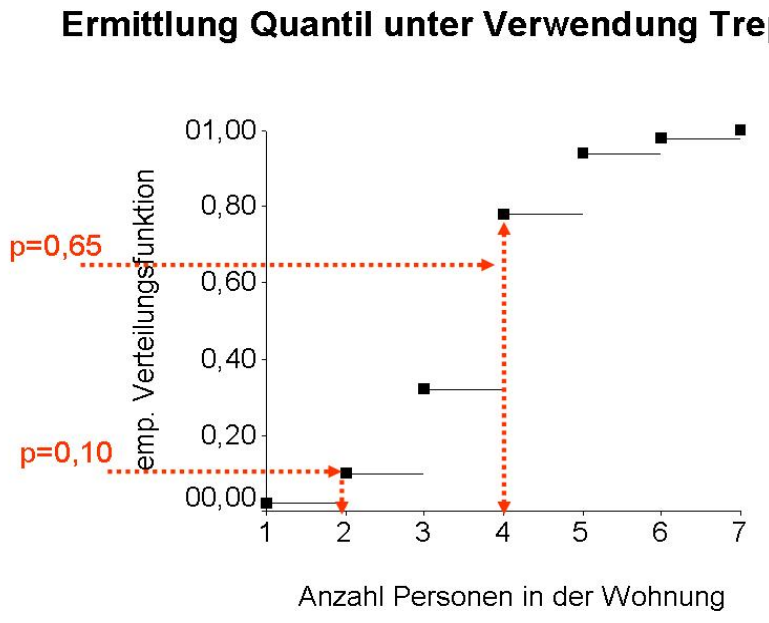
Graph der empirischen Verteilungsfunktion



empirische Verteilungsfunktion im analytischen Ausdruck

$$F(x) = \begin{cases} 0 & \text{falls} \quad x < 1 \\ 0{,}02 & \text{falls} \quad 1 \le x < 2 \\ 0{,}10 & \text{falls} \quad 2 \le x < 3 \\ 0{,}32 & \text{falls} \quad 3 \le x < 4 \\ 0{,}78 & \text{falls} \quad 4 \le x < 5 \\ 0{,}94 & \text{falls} \quad 5 \le x < 6 \\ 0{,}98 & \text{falls} \quad 6 \le x < 7 \\ 1{,}00 & \text{falls} \quad x \ge 7 \end{cases}$$

Квантиль – определение из эмпирической функции распределения



Дополнительные квантили: 0,20-квантиль

Квантиль – определение для интервалов

Исходная точка: интервал $I_l = (c_{l-1}, c_l]$

Квантиль: $x_p \in I_l \;\Rightarrow\; c_{l-1} < x_p \leq c_l$

$$x_p \approx c_{l-1} + \frac{p - F_{l-1}}{p_l^D}$$

где $F_{l-1} < p < F_l$

Пример из лекции 2

Häufigkeitsverteilung, tabellarisch

j	x_j^u	x_j^o	n_j	p_j
1	0	2	58	0,58
2	2	4	22	0,22
3	4	6	16	0,16
4	6	10	3	0,03
m=5	10	14	1	0,01
Σ			n=100	1,00



Klassenunter- bzw. -oberarenze

Δ_j	p_j^D
2	0,2900
2	0,1100
2	0,0800
4	0,0075
4	0,0025

Вычисление 0,7-квантиля:

$$
\begin{aligned}
g &= n \cdot p = 100 \cdot 0{,}7 = 70 \\
x_{\mathbf{0,7}} &\in (2, 4] = (c_{\mathbf{1}}, c_{\mathbf{2}}] \\
F_{\mathbf{1}} &= 0{,}58 \\
F_{\mathbf{2}} &= 0{,}80 \\
x_p &\approx c_{l-\mathbf{1}} + \frac{p - F_{l-\mathbf{1}}}{p_l^D} = 2 + \frac{0{,}7 - 0{,}58}{0{,}11} = 3{,}09
\end{aligned}
$$

Пример из лекции 2

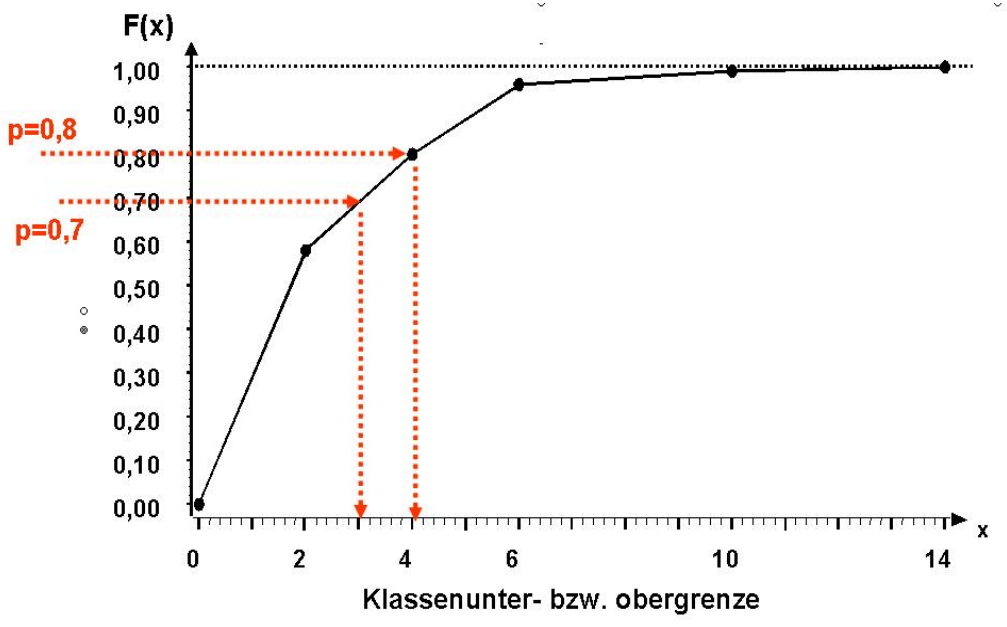
Интерпретация:

$x_{0,7} = 3{,}09$ мин $= 3$ мин $+ 5{,}4$ сек

1. Не менее 70% всех телефонных разговоров не превышают продолжительности 3,09 мин.
2. Максимальная продолжительность 70% всех телефонных разговоров составляет 3,09 мин.
3. Не более 30% всех телефонных разговоров длиннее 3,09 мин.

Примечание: сравнение с 70-процентным квантилем из отдельных значений ($x_{0,7} = 3{,}11$)

Квантиль для интервалов – из эмпирической функции распределения



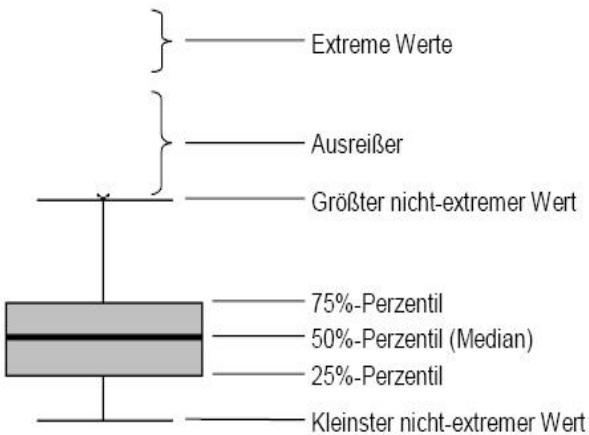
Дополнительные квантили: 0,05-квантиль; 0,02-квантиль; 0,19-квантиль

Специальные квантили

p	Название квантиля
0,01	Перцентиль
0,10	Дециль (нижний)
0,25	Квартиль (нижний)
0,5	Медиана, квартиль (средний)
0,75	Квартиль (верхний)
0,90	Дециль (верхний)
0,99	Перцентиль

Диаграмма "ящик с усами"

Боксплот (Box-and-Whisker-Plot) описывает распределение значений. Он показывает квартили, экстремальные значения и выбросы, а также наибольшее и наименьшее не экстремальное значение. Экстремально большие значения или выбросы вверх характеризуются тем, что их расстояние до 75-го перцентиля больше 1,5-кратного расстояния между 75-м и 25-м перцентилями. Экстремально малые значения и выбросы оцениваются аналогично по расстоянию до 25-го перцентиля.



Введение в SPSS – темы

Статистические темы: дескриптивная статистика

- Частоты – продолжение
- Диаграммы (боксплоты) – продолжение

Мода, медиана, квантили – пример

с Excel:

Пример 1 (из первичного списка): BeispielW95.xls
Пример 2 (из классифицированных данных): Zinssaetze 2007_1.xls:

Задание: определить моду, медиану, квантили

с SPSS:

Пример 1: World95.sav
Пример 2 (из классифицированных данных): Zinssaetze 2007_1.xls:

Задание: определить моду, медиану, квантили



2.2.1 Меры положения – Содержание

- 2.2.1.1 Введение
- 2.2.1.2 Мода
- 2.2.1.3 Медиана
- 2.2.1.4 Квантили
- **2.2.1.5 Пример: Историческое моделирование**
- 2.2.1.6 Среднее арифметическое

Определения 1

Пусть задан портфель, рыночную стоимость которого можно описать как функцию от рыночных данных (например, обменных курсов или процентных ставок).

$P(t)$ – рыночная стоимость портфеля в момент t
$P(t_0)$ – рыночная стоимость портфеля в начальный момент t_0

Допущение:

состав портфеля не изменяется
в течение периода $t - t_0 > 0$

$\Delta P(t) = P(t) - P(t_0)$ – изменение стоимости портфеля

(прибыли и убытки – PnL)

Определения 2

Обозначения:

$Y(t)$ – вектор рыночных данных в момент t
$Y_i(t)$ – i-я компонента вектора рыночных данных
$P(Y(t))$ – ценовая функция (стоимость портфеля в момент t)

Изменение стоимости портфеля (PnL) можно представить так:

$$\Delta P(t) = P(Y(t)) - P(Y(t_0))$$

Value at Risk (VaR)

Цель оценки рыночного риска – получить суждение о потенциальных потерях стоимости портфеля в течение фиксированного периода удержания t.

Потенциальные потери стоимости называются Value at Risk (VaR).

Допущение: располагаем статистическим материалом об изменениях стоимости портфеля. Сначала изменения стоимости портфеля

упорядочиваются по величине: $\Delta P(t) < \Delta P(t+1)$ Вычисляется

$\alpha \cdot 100\%$-квантиль:

$x_\alpha = min(\Delta P(t) \,|F(\Delta P(t)) \geq \alpha)$

Математически $VaR(t, \alpha)$ является α-квантилем распределения $\Delta P(t)$.

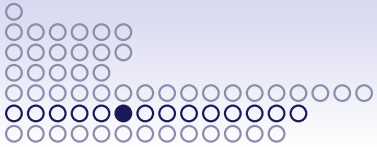
Value at Risk (VaR)

Интерпретация

1. Максимальное значение, достигаемое $\alpha \cdot 100\%$ всех случаев изменений стоимости портфеля ($\Delta P(t)$), равно x_α.
2. В не менее чем $\alpha \cdot 100\%$ всех случаев изменений стоимости портфеля ($\Delta P(t)$) значение x_α не превышается.
3. Не более чем в $(1 - \alpha) \cdot 100\%$ всех случаев значение x_α превышается.

Примечание: как правило, первые случаи изменений стоимости портфеля ($\Delta P(t)$) отрицательны, поэтому говорят о потерях стоимости.

1. Минимальная потеря стоимости, достигаемая $\alpha \cdot 100\%$ всех случаев изменений стоимости портфеля, равна x_α.
2. Не более чем в $(1 - \alpha) \cdot 100\%$ всех случаев потеря стоимости x_α не достигается.



Содержание

- Идея исторического моделирования
- Метод
- Пример

Идея исторического моделирования 1

Дано:

- Текущие рыночные данные – $Y_i(0)$, $i = 1, 2, ..., n$
- Исторические рыночные данные – $Y_i(t)$, $t = -1, -2, ..., -T$
- Текущая рыночная стоимость портфеля $P(0) = P(Y_1(0), Y_2(0), ..., Y_n(0))$

Вместо рыночных данных используем так называемую **логарифмическую доходность** (log-returns):

$$R_i(0) = ln(\frac{Y_i(0)}{Y_i(-1)})$$

в общем:

$$R_i(t) = ln(\frac{Y_i(t)}{Y_i(t-1)}),\ t = 0, -1, ..., -T+1$$

Примечание: обычно $T = 250$

Идея исторического моделирования 2

Дано:

- Текущие рыночные данные (лог. доходность) – $R_i(0)$, $i = 1, 2, ..., n$
- Исторические рыночные данные (лог. доходность) – $R_i(t)$, $t = -1, -2, ..., -250$
- Текущая рыночная стоимость портфеля $P(0) = P(R_1(0), R_2(0), ..., R_n(0))$

Оцениваем текущий портфель с историческими рыночными данными (лог. доходность):

$$P(t) = P(R_1(t), R_2(t), ..., R_n(t)),\ t = -1, -2, ..., -250$$

Вычисляем изменения стоимости портфеля (текущие vs. исторические данные)

$$\begin{aligned} \Delta P(t) &= P(t) - P(0) \\ &= P(R_1(t), ..., R_n(t)) - P(R_1(0), ..., R_n(0)) \end{aligned}$$

$\Rightarrow$ 250 изменений стоимости портфеля:

$\Delta P(-1), \Delta P(-2), ..., \Delta P(-250) \Rightarrow$ дискретное распределение

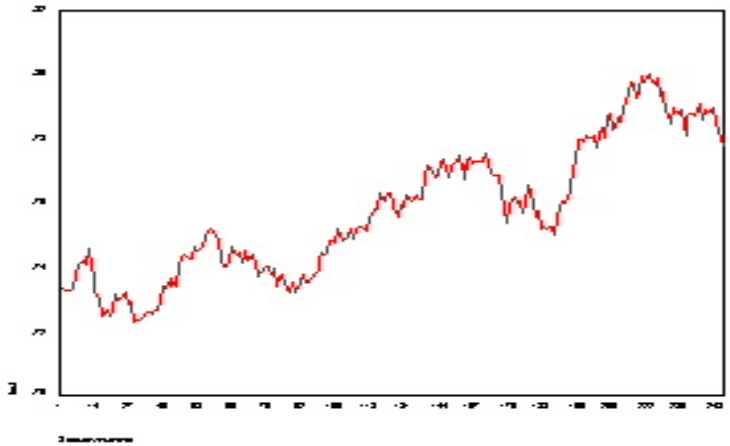
Идея исторического моделирования 3

Определение квантиля дискретного распределения

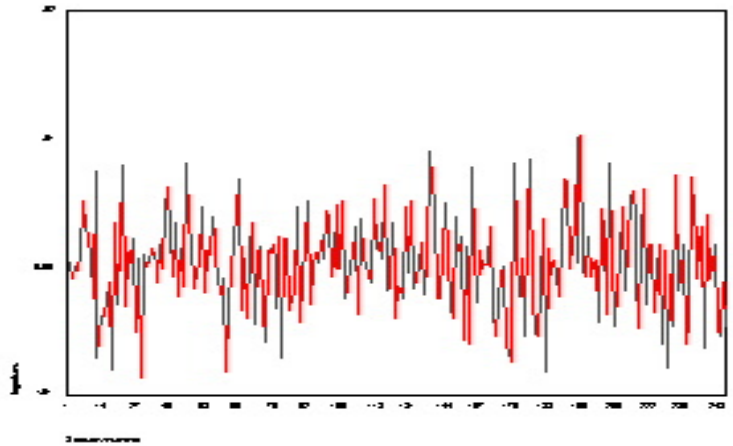
- Упорядочить изменения стоимости по величине $P(1) \leq P(2) \leq ... \leq P(250)$ (новая нумерация!)
- Вычислить квантиль дискретного распределения при заданном уровне доверия α
- $\alpha = 0{,}01$ при $n = 250 \Rightarrow$ среднее второго и третьего наименьших изменений $\frac{1}{2}(P(2) + P(3))$
- $\alpha = 0{,}10$ при $n = 250 \Rightarrow$ 25-е наименьшее изменение $(P(25))$
- $\alpha = 0{,}05$ при $n = 250 \Rightarrow$ среднее 12-го и 13-го наименьших изменений $\frac{1}{2}(P(12) + P(13))$

Пример: VaR FX-кассовой позиции – историческое моделирование

Позиция $N = 1\,000\,000$ USD, базовая валюта EUR
Период удержания $t = 1$
Уровень доверия $\alpha = 0{,}01$
Обменный курс на 31.08.2007: $Y(t_0) = 0{,}73338$



Исходные данные $Y(t)$



Логарифмическая доходность

Обменный курс доллар – евро (04.09.2006–31.08.2007)
$\Rightarrow$ 251 наблюдение

(Обратная временная ось)

Пример: VaR FX-кассовой позиции – ист. мод.

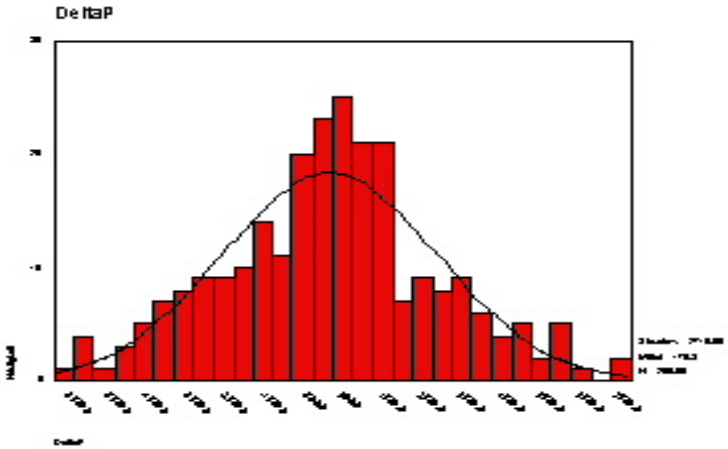
t	Дата	$Y_i(t+1)$	$R_{i+\mathbf{1}}(t)$	$\widetilde{Y_i(t+1)}$	$\Delta P(t+1)$
-1	2007-08-31	0.7334	-0.000388	0.7331	-284.5
-2	2007-08-30	0.7337	0.000925	0.7341	679.0
-3	2007-08-29	0.7330	-0.000005	0.7334	-3.5
-4	2007-08-28	0.7330	0.000259	0.7336	190.2
...	...	...	...	...	...

$$
\begin{aligned}
R_i(t) &= ln(\frac{Y_i(t)}{Y_i(t-1)}),\ t = -1, -2, \ldots, -250 \\
\widetilde{Y_i(t)} &= e^{R_i(t+\mathbf{1})} Y_i(0),\ t = -1, -2, \ldots, -250 \\
\Delta P(t+1) &= N(\widetilde{Y_i(t)} - Y_i(0)),\ t = -1, -2, \ldots, -250
\end{aligned}
$$

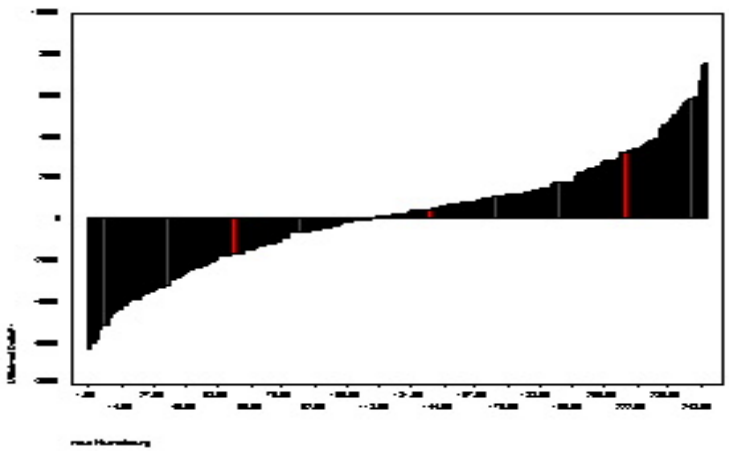
например, для $t = -1$

$$
\begin{aligned}
R_i(0) &= ln(\frac{0.7334}{0.7337}) = -0.000388 \\
\widetilde{Y_i(0)} &= e^{-\mathbf{0.000388}} \cdot 0.7334 = 0.7331 \\
\Delta P(0) &= 1\,000\,000 \cdot (0.7331 - 0.7334) = -284.5
\end{aligned}
$$

VaR FX-кассовой позиции – ист. мод.



$\Delta P(t)$



Упорядоченные $\Delta P(t)$

Пример: VaR FX-кассовой позиции – ист. мод.

Дата	$\Delta P(t)$	t	Новая нум.
2006-11-22	-7535.5	-194	1
2006-11-24	-7393.4	-192	2
2007-02-14	-6575.9	-137	3
2006-12-19	-6072.0	-175	4
2006-11-07	-5924.5	-205	5
2007-07-02	-5896.3	-44	6
...	...	...	...

Внимание: использовать новую нумерацию

$$VaR(\alpha = 0{,}01) = \frac{1}{2}(P(2) + P(3)) = \frac{1}{2}(-7393{,}4 - 6575{,}9) = -6984{,}65$$

Замечание 1: учёт антитетики

При формировании сценариев с учётом антитетики к каждому движению рынка, наблюдавшемуся в историческом периоде, в матрицу сценариев добавляется также точно противоположное движение рынка. Таким образом учитывается, что даже в случае однонаправленной положительной рыночной тенденции или противоположной чувствительности портфеля за период наблюдения рассчитывается адекватный риск.

$$\widetilde{\Delta Y_i(t)} = e^{R_i(t)} Y_i(0),\ t = -1, -2, ..., -250$$

и

$$\widehat{\Delta Y_i(t)} = e^{-R_i(t)} Y_i(0),\ t = -1, -2, ..., -250$$



2.2.1 Меры положения – Содержание

- 2.2.1.1 Введение
- 2.2.1.2 Мода
- 2.2.1.3 Медиана
- 2.2.1.4 Квантили
- 2.2.1.5 Пример: Историческое моделирование
- **2.2.1.6 Среднее арифметическое**

Среднее арифметическое – определения 1

Функция расстояния: $d(x_i, a) = (x_i - a)^2$

Требование: $f(a) = \sum_{i=1}^{n} d(x_i, a) = \sum_{i=1}^{n} (x_i - a)^2 =$ минимум

Задача на экстремум для функции одной переменной.

$$\begin{aligned} \frac{\partial f(a)}{\partial a} &= 0 \\ &= -2\sum_{i=1}^{n}(x_i - a) = -2(\sum_{i=1}^{n} x_i - na) = 0 \quad \Rightarrow \\ a &= \frac{\sum_{i=1}^{n} x_i}{n} \end{aligned}$$



Среднее арифметическое – определения 2

Определение:

Если X – кардинально масштабированный признак статистической совокупности объёма n, то значение a, которое получается при равномерном распределении суммы всех наблюдаемых значений признака x_i $(i = 1, 2, ..., n)$ по всем носителям признака, называется **средним арифметическим**.

Обозначение: $\overline{x}$

$$\overline{x} = a$$

Среднее арифметическое – вычисления

- Из первичного списка:

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

- Из таблицы частот (несовпадающие признаки):

$$\overline{x} = \frac{\sum_{l=1}^{m} \xi_l \cdot n_l}{n} = \sum_{l=1}^{m} \xi_l \cdot p_l$$

- Из таблицы частот (интервалы):

Среднее арифметическое – вычисления

Из таблицы частот (интервалы):

$$\overline{x} \approx \frac{\sum_{l=1}^{r} \widetilde{x}_l \cdot n_l}{n} = \sum_{l=1}^{r} \widetilde{x}_l \cdot p_l$$

где $\widetilde{x}_l = \frac{1}{2}(c_{l-1} + c_l)$

Среднее арифметическое – примеры

1. Из первичного списка:
   Пример из лекции 1 (Лист Excel)
2. Из таблицы частот (несовпадающие признаки):
   Пример из лекции 1 (Лист Excel)
3. Из таблицы частот (интервалы):
   Пример из лекции 2 (Лист Excel)

Среднее арифметическое – свойства – I

- Свойство нуля / центра тяжести:

$$\sum_{i=1}^{n}(x_i - \overline{x}) = 0$$

$$\sum_{l=1}^{m}(x_l - \overline{x})n_l = 0$$

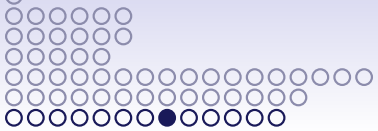
- Свойство экстраполяции:

$$\sum_{i=1}^{n} x_i = \overline{x} \cdot n$$

- Среднее арифметическое для линейно преобразованного набора данных:

$$y_i = a + bx_i \Rightarrow \overline{y} = a + b\overline{x}$$

- Чувствительность к выбросам



Объединённый набор данных

Определение: Набор данных подразделяется на несовпадающие группы $D_1, D_2, ..., D_r$. Для каждой группы статистические характеристики вычислены и известны. Такой набор данных называется **объединённым набором данных** (пулом данных).

Задача: Вычислить статистические характеристики для объединённого набора данных $(D_1 \cup D_2 \cup ... \cup D_r)$ с помощью характеристик несовпадающих групп $D_1, D_2, ..., D_r$.

Среднее арифметическое для объединённого набора данных

Если наблюдаемые значения заданы в несовпадающих группах $D_1, D_2, ..., D_r$ и для каждой группы известно среднее арифметическое $\overline{x_p}$, то среднее арифметическое для объединённого набора данных $(D_1 \cup D_2 \cup ... \cup D_r)$ вычисляется по формуле, где n_p – абсолютные частоты значений p-й группы:

$$\overline{x} = \frac{\sum_{p=1}^{r} \overline{x_p} \cdot n_p}{n}$$

$$\text{где } n = \sum_{p=1}^{r} n_p$$

Среднее арифметическое (объединённый набор) – примеры

Набор данных: Пример из лекции 2

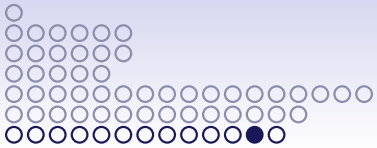
j	c_u	c_o	n_j	p_j	arithmetisches Mittel	empirische Varianz
1	0	2	58	0.58	0.92	0.3508
2	2	4	22	0.22	2.9	0.3694
3	4	6	16	0.16	4.63	0.2578
4	6	10	3	0.03	7.98	0.038
5	10	14	1	0.01	13.13	0
			100	1		

Примечание:
Средние значения и эмпирические дисперсии вычислены из первичного списка для соответствующих групп.

$$\overline{x} = \frac{\sum_{p=1}^{r} \overline{x_p} \cdot n_p}{n} = 2{,}2821$$

Средние значения – специальные случаи

- Среднее для временных рядов (среднегодовой темп роста) – геометрическое среднее
- Гармоническое среднее (статистический признак является относительным показателем)
- Взвешенное среднее арифметическое



Введение в SPSS – темы

Формальная работа с SPSS

- Работа с файлами и преобразование файлов – продолжение (группировка в SPSS)

Статистические темы: дескриптивная статистика

- Частоты – продолжение

Средние значения – пример

с Excel:

Пример 1 (из первичного списка): BeispielW95.xls
Пример 2 (из классифицированных данных): Zinssaetze 2007_1.xls
Пример 3 (объединённый набор): Zinssaetze 2007_4.xls

Задание: определить средние значения

с SPSS:

Пример 1 (с группировкой в SPSS): World95.sav.xls
Пример 2 (из классифицированных данных): Zinssaetze 2007_1.xls

Задание: определить средние значения









Числовые характеристики распределения

- 2.2.1 Меры положения (меры центральной тенденции)
- **2.2.2 Меры рассеяния**
- 2.2.3 Меры формы
- 2.2.4 Линейные преобразования

2.2.2 Меры рассеяния – Содержание

- **2.2.2.1 Меры диспаритета**
- 2.2.2.2 Метрические меры рассеяния

Меры рассеяния – намерение

Цель:
Предоставить информацию, количественно оценивающую концентрацию наблюдений вокруг меры положения. В зависимости от используемых шкал применяются следующие характеристики:

- Номинальная шкала $\Rightarrow$ номинальная мера диспаритета
- Порядковая шкала $\Rightarrow$ порядковая мера диспаритета
- Кардинальная шкала (метрическая шкала)
  - Размах
  - Центральный квантильный интервал
  - (Эмпирическая) дисперсия
  - (Эмпирическое) стандартное отклонение
  - Коэффициент вариации

Номинальная мера диспаритета

Исходная точка: относительное распределение частот **номинального** признака:

$$f : \xi_l \to p_l = f(\xi_l) = p(\xi_l),\ l = 1, 2, ..., m$$

Определение: **номинальная мера диспаритета** d_N:

$$d_N = \sqrt{1 - \frac{m}{m-1}(1 - \sum_{l=1}^{m} p_l^2)} \text{ при } 0 \leq d_N < 1$$

Часто также: нормированная мера рассеяния по ХЕРФИНДАЛЮ

$$d_H = \frac{m}{m-1}(1 - \sum_{l=1}^{m} p_l^2) \text{ при } 0 \leq d_H < 1$$

Интерпретация:

$d_N \approx 0$ или $d_H \approx 1$ $\Rightarrow$ равномерное распределение признака.

$d_N \approx 1$ или $d_H \approx 0$ $\Rightarrow$ почти вырожденное распределение.

Номинальная мера диспаритета – интерпретация

Допущение: равномерное распределение $\Rightarrow \quad p_l = \frac{1}{m}$

$$\begin{aligned} \sum_{l=1}^{m} p_l^2 &= \frac{1}{m} \\ d_N &= \sqrt{1 - \frac{m}{m-1} \cdot (1 - \frac{1}{m})} = 0 \end{aligned}$$

Допущение: вырожденное распределение $\Rightarrow \quad p_1 = 1;\ p_2 = \ldots = p_m = 0$

$$\begin{aligned} \sum_{l=1}^{m} p_l^2 &= 1 \\ d_N &= \sqrt{1 - \frac{m}{m-1} \cdot (1 - 1)} = 1 \end{aligned}$$

Номинальная мера диспаритета – Пример из лекции 4

Регион или экономическая группа (1995)

	Несовп. знач. признака	Абс. частоты	Отн. частоты
j	ξ_j	n_j	p_j
1	OECD	21	19.3
2	Восточная Европа	14	12.8
3	Тихоокеанская Азия	17	15.6
4	Африка	19	17.4
5	Ближний Восток	17	15.6
6	Латинская Америка	21	19.3
Сумма	Итого	109	100.0

$$d_N = \sqrt{1 - \frac{m}{m-1}(1 - \sum_{l=1}^{m} p_l^2)} = \sqrt{1 - \frac{6}{5}(1 - 0{,}169767)} = 0{,}060994$$

Интерпретация:

$d_N \approx 0 \Rightarrow$ **приближённое равномерное распределение признака.**

Номинальная мера диспаритета – Пример из лекции 4

Регион или экономическая группа (1995)

	Несовп. знач. признака	Абс. частоты	Отн. частоты
j	ξ_j	n_j	p_j
1	OECD	21	19.3
2	Восточная Европа	14	12.8
3	Тихоокеанская Азия	17	15.6
4	Африка	19	17.4
5	Ближний Восток	17	15.6
6	Латинская Америка	21	19.3
Сумма	Итого	109	100.0

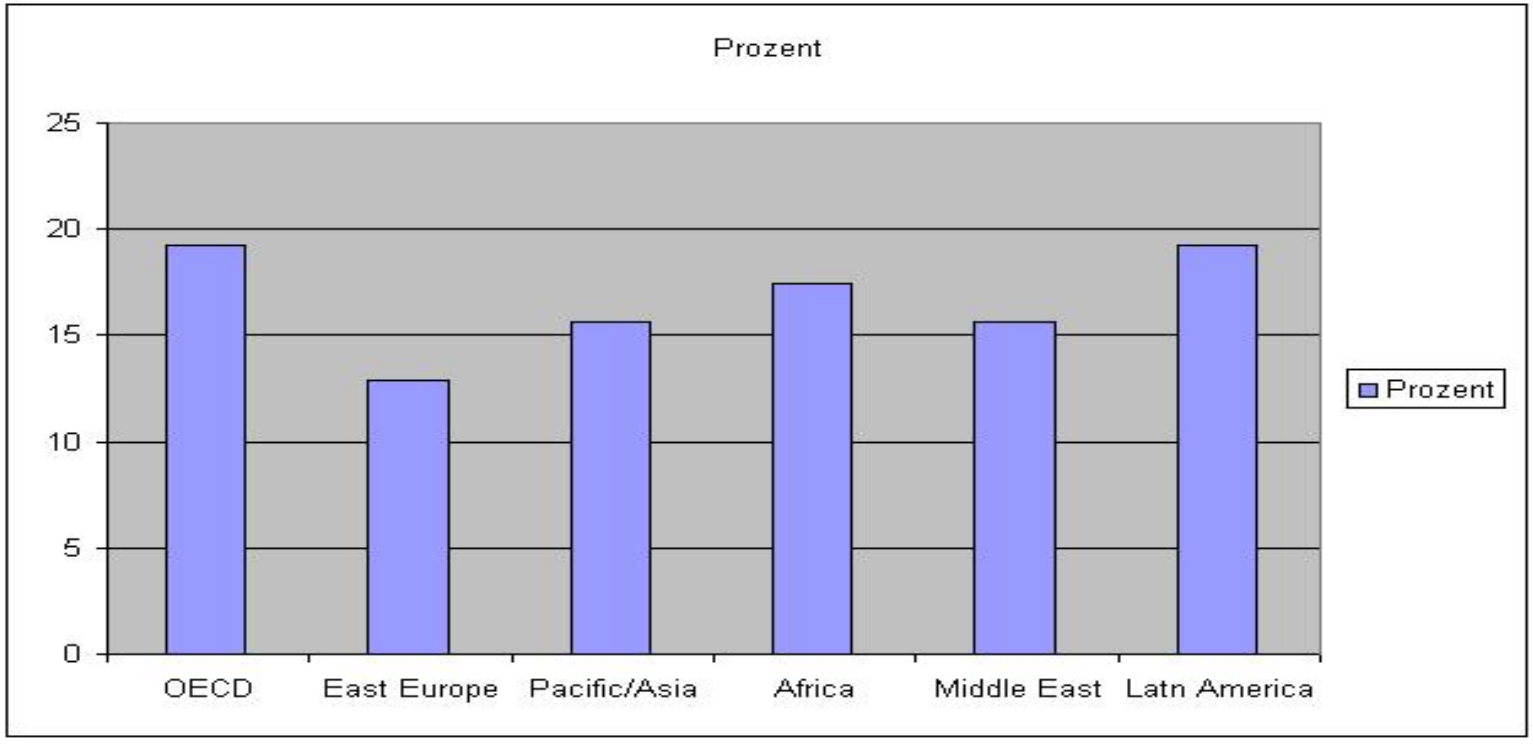
$$d_H = \frac{m}{m-1}(1 - \sum_{l=1}^{m} p_l^2) = \frac{6}{5}(1 - 0{,}169767) = 0{,}996280$$

Интерпретация:

$d_H \approx 1 \Rightarrow$ **приближённое равномерное распределение признака.**

Номинальная мера диспаритета – Пример из лекции 4

Набор данных: Пример из лекции 4



Примечание:

$d_N \approx 0 \Rightarrow$ **приближённое равномерное распределение признака.**

Порядковая мера диспаритета

Исходная точка: абсолютное кумулятивное распределение **порядкового** признака:

$$f : \xi_l \to F_l = f(\xi_l) = F(X \leq \xi_l)$$

Определение: **порядковая мера диспаритета** d_O:

$$d_O = 1 - \frac{6m}{m^2 - 1} \sum_{l=1}^{m-1} F_l(1 - F_l) \text{ при } 0 \leq d_O < 1$$

Интерпретация:

$d_O = 0 \Rightarrow$ равномерное распределение признака.

$d_O \approx 1 \Rightarrow$ почти вырожденное распределение.

Порядковая мера диспаритета – Пример из лекции 5

Оценки за экзамен

	Несовп. знач. признака	Абс. частоты	Отн. частоты	Отн. кум. частоты
j	ξ_j	n_j	p_j	F_j
1	отлично	5	11.4	11.4
2	хорошо	20	45.5	56.9
3	удовлетворительно	10	22.7	79.6
4	достаточно	6	13.6	93.2
5	неудовлетворительно	4	6.8	100.0
Сумма	Итого	45	100.0	

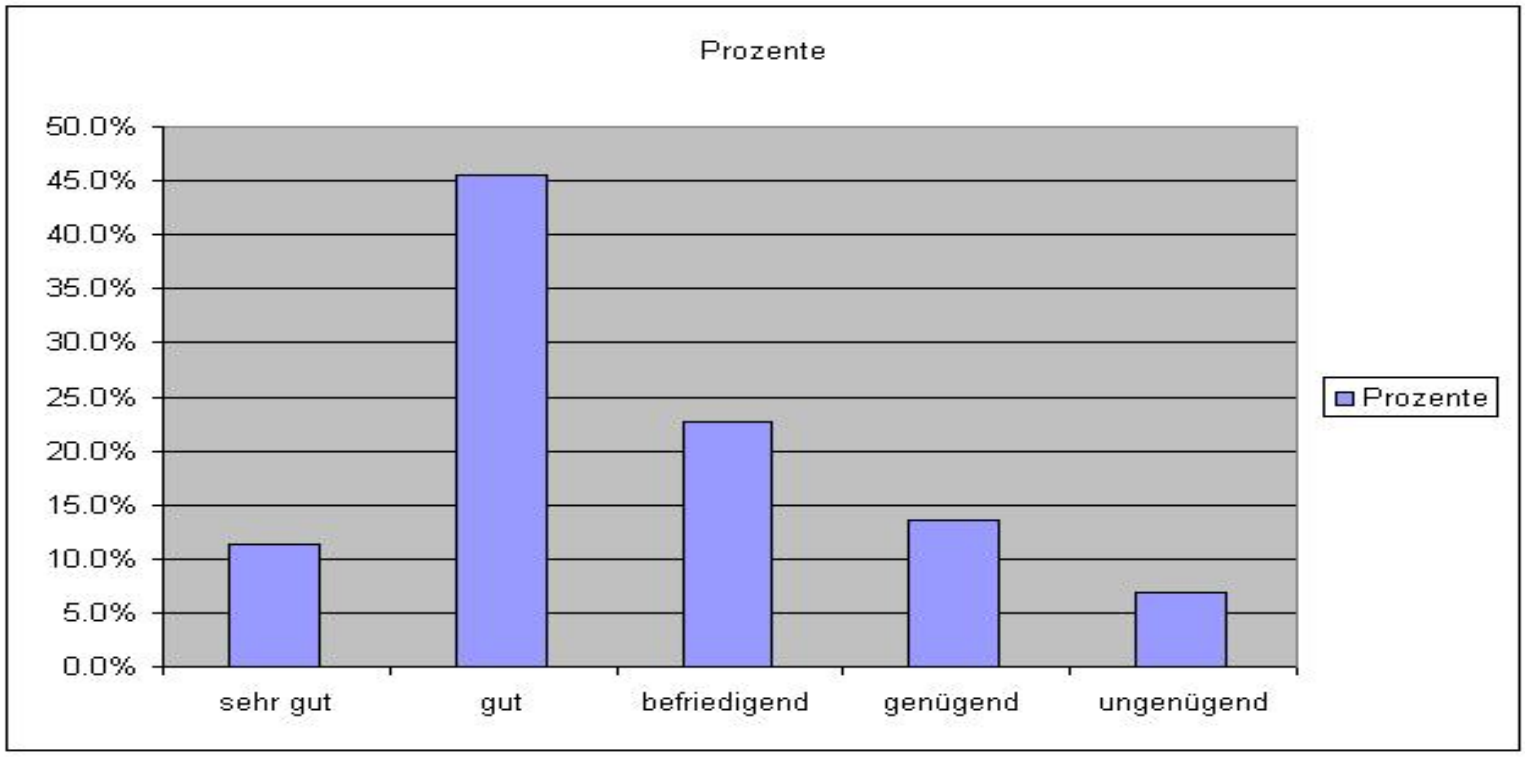
$$d_O = 1 - \frac{6m}{m^2 - 1} \sum_{l=1}^{m-1} F_l(1 - F_l) = 1 - \frac{6 \cdot 5}{24} \sum_{l=1}^{4} F_l(1 - F_l) = 0{,}2846$$

Интерпретация:

$d_O > 0 \Rightarrow$ **распределение далеко от равномерного.**

Порядковая мера диспаритета – Пример из лекции 5

Набор данных: Пример из лекции 4



Примечание:

$d_O > 0 \Rightarrow$ **распределение далеко от равномерного.**









2.2.2 Меры рассеяния – Содержание

- 2.2.2.1 Меры диспаритета
- **2.2.2.2 Метрические меры рассеяния**

Размах

Пусть задан ранговый ряд порядкового набора данных:

$$x_1 \leq x_2 \leq \ ..., \leq x_n$$

Определение: **Размах R**:

$$R = X_{MAX} - X_{MIN} = x_n - x_1$$

Определение: **Размах R** с группировкой по интервалам:

$$R = c_r - c_0$$

Центральный квантильный интервал

Пусть X – кардинально масштабированный, упорядоченный по возрастанию признак статистической совокупности:

$$x_1 \leq x_2 \leq \ ..., \leq x_n$$

Тогда разность между квантилем $(1 + p)/2$ и квантилем $(1 - p)/2$ при заданном p, $0 < p < 1$, называется центральным квантильным интервалом (межквантильным интервалом) порядка p и обозначается Q_p.

Специальный центральный квантильный интервал:

$$Q_p = x_{\frac{1+p}{2}} - x_{\frac{1-p}{2}}$$

Центральный квантильный интервал порядка $p = 0{,}5$ (межквартильный интервал):

$$Q_{0,5} = x_{0,75} - x_{0,25}$$

Эмпирическая дисперсия – формулы вычисления

Если X – кардинально масштабированный признак статистической совокупности, то среднее арифметическое квадратов отклонений значений признака от их среднего арифметического называется эмпирической дисперсией d_x^2:

- Из первичного списка (наблюдения x_i, $i = 1, 2, ..., n$):

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

- Из таблицы частот (несовпадающие признаки ξ_l, относительные частоты p_l, $l = 1, ..., m$):

$$d_x^2 = \sum_{l=1}^{m} (\xi_l - \overline{x})^2 \cdot p_l$$

Эмпирическая дисперсия – формулы вычисления

- Из таблицы частот с группировкой (середина интервала x_l^*, относит. частоты p_l, $l = 1, ..., r$):

$$d_x^2 = \sum_{l=1}^{r} (x_l^* - \overline{x})^2 \cdot p_l$$

- Для объединённых данных (среднее $\overline{x_l}$, дисперсия d_l^2, отн. частоты p_l группы l, $l = 1, ..., r$):

$$d_x^2 = \sum_{l=1}^{r} d_l^2 \cdot p_l + \sum_{l=1}^{r} (\overline{x_l} - \overline{x})^2 \cdot p_l = d_a^2 + d_b^2$$

d_a^2 – дисперсия внутри групп
d_b^2 – дисперсия между группами

Эмпирическая дисперсия для объединённых данных – вывод

$$
\begin{aligned}
d_x^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2 = \frac{1}{n}\sum_{l=1}^{m}\sum_{i=1}^{n_l}(x_{il}-\overline{x})^2 = \\
&= \frac{1}{n}\sum_{l=1}^{m}\sum_{i=1}^{n_l}[(x_{il}-\overline{x_l})+(\overline{x_l}-\overline{x})]^2 = \\
&= \frac{1}{n}\sum_{l=1}^{m}\sum_{i=1}^{n_l}(x_{il}-\overline{x_l})^2 + \frac{1}{n}\sum_{l=1}^{m}\sum_{i=1}^{n_l}(\overline{x_l}-\overline{x})^2 \\
&= \frac{1}{n}\sum_{l=1}^{m} n_l(x_{il}-\overline{x_l})^2 + \frac{1}{n}\sum_{l=1}^{m} n_l(\overline{x_l}-\overline{x})^2 \\
&= \sum_{l=1}^{m} d_l^2\cdot p_l + \sum_{l=1}^{m}(\overline{x_l}-\overline{x})^2\cdot p_l = d_a^2 + d_b^2
\end{aligned}
$$

Эмпирическое стандартное отклонение – определение

(Положительный) квадратный корень из эмпирической дисперсии кардинально масштабированного признака X называется эмпирическим стандартным отклонением (d_x).

$$d_x = \sqrt{d_x^2}$$

Эмпирическое стандартное отклонение – интерпретация

- В интервале $[\overline{x} - d_x, \overline{x} + d_x]$ лежит не менее 50% всех наблюдений.
- При нормально распределённых наблюдениях в интервале $[\overline{x} - d_x, \overline{x} + d_x]$ лежит около 68% всех наблюдений.
- При нормально распределённых наблюдениях в интервале $[\overline{x} - 2d_x, \overline{x} + 2d_x]$ лежит около 95% всех наблюдений.
- При нормально распределённых наблюдениях в интервале $[\overline{x} - 3d_x, \overline{x} + 3d_x]$ лежит около 99% всех наблюдений.

Коэффициент вариации

Отношение эмпирического стандартного отклонения d_x к среднему арифметическому $\overline{x}$ называется **коэффициентом вариации** (V_x):

$$V_x = \frac{d_x}{\overline{x}}$$

$$\text{Условие :} \quad \overline{x} > 0 \text{ и } x_i > 0 \ \forall i = 1, 2, ..., n$$

Интерпретация: коэффициент вариации больше 0,5 является признаком того, что среднее арифметическое $\overline{x}$ из-за очень большого рассеяния d_x не является подходящим представителем отдельных значений x_i.

Метрические меры рассеяния – пример 1

Задан упорядоченный первичный список (телефонные разговоры):

$$x_1 \leq x_2 \leq \ ..., \leq x_n$$

$$\begin{aligned}
\overline{x} &= \frac{\sum_{i=1}^{n} x_i}{n} \\
R &= X_{MAX} - X_{MIN} = X_n - X_1 \\
Q_p &= x_{\frac{1+p}{2}} - x_{\frac{1-p}{2}} \\
d_x^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \\
d_x &= \sqrt{d_x^2} \\
V_x &= \frac{d_x}{\overline{x}}
\end{aligned}$$

Меры рассеяния – Пример из лекции 1

aufsteigend sortierte Urliste:

0,04	0,09	0,10	0,10	0,11
0,12	0,14	0,16	0,22	0,23
0,26	0,27	0,27	0,34	0,38
0,39	0,44	0,46	0,48	0,51
0,51	0,55	0,63	0,66	0,76
0,77	0,84	0,87	0,87	0,97
0,97	0,97	1,03	1,05	1,06
1,07	1,15	1,19	1,20	1,26
1,29	1,33	1,33	1,42	1,54
1,57	1,57	1,62	1,68	1,70
1,72	1,74	1,74	1,78	1,79
1,86	1,93	1,99	2,01	2,02
2,13	2,14	2,15	2,28	2,38
2,58	2,59	2,82	2,92	3,11
3,11	3,13	3,15	3,23	3,47
3,51	3,60	3,64	3,86	3,90
4,11	4,14	4,23	4,28	4,32
4,32	4,41	4,41	4,52	4,59
4,60	4,67	4,68	5,39	5,42
5,98	7,71	8,05	8,17	13,13

Метрические меры рассеяния – пример 2

Задана таблица частот с несовпадающими значениями признака (напр., 50 квартир):

$$\xi_1 \leq \xi_2 \leq \ldots, \leq \xi_m$$

$$\begin{aligned}
\overline{x} &= \sum_{i=1}^{m} \xi_i \cdot p_i \\
R &= \xi_{MAX} - \xi_{MIN} \\
Q_p &= x_{\frac{1+p}{2}} - x_{\frac{1-p}{2}} \\
d_x^2 &= \sum_{i=1}^{m} (\xi_i - \overline{x})^2 \cdot p_i \\
d_x &= \sqrt{d_x^2} \\
V_x &= \frac{d_x}{\overline{x}}
\end{aligned}$$

Меры рассеяния – Пример из лекции 1

Задание: определить 0,05-квантиль ($p = 0{,}05$)

	Несовп. знач. признака	Абс. частоты	Отн. частоты
j	ξ_j	n_j	p_j
1	1	1	0.02
2	2	4	0.08
3	3	11	0.22
4	4	23	0.46
5	5	8	0.16
6	6	2	0.04
7	7	1	0.02
Сумма		50	1

Метрические меры рассеяния – пример 3

Задана классификация (интервалы, телефонные разговоры)

$$I_1, I_2, ..., I_r = [c_0, c_1], (c_1, c_2], ..., (c_{r-1}, c_r]$$

$$\begin{aligned}
\overline{x} &\approx \sum_{l=1}^{r} \widetilde{x}_l \cdot p_l \\
R &= c_r - c_0 \\
Q_p &= x_{\frac{1+p}{2}} - x_{\frac{1-p}{2}} \\
d_x^2 &= \sum_{l=1}^{r} (\widetilde{x}_l - \overline{x})^2 \cdot p_l \\
d_x &= \sqrt{d_x^2} \\
V_x &= \frac{d_x}{\overline{x}}
\end{aligned}$$

Меры рассеяния – Пример из лекции 2

Häufigkeitsverteilung, tabellarisch

j	x_j^u	x_j^o	n_j	p_j
1	0	2	58	0,58
2	2	4	22	0,22
3	4	6	16	0,16
4	6	10	3	0,03
m=5	10	14	1	0,01
Σ			n=100	1,00



Klassenunter- bzw. -oberarenze

Δ_j	p_j^D
2	0,2900
2	0,1100
2	0,0800
4	0,0075
4	0,0025

Метрические меры рассеяния – пример 4

Задан объединённый набор данных (телефонные разговоры)

Несовпадающие группы $D_1, D_2, ..., D_r$ с $(D_1 \cup D_2 \cup ... \cup D_r)$

$$
\begin{aligned}
\overline{x} &= \frac{\sum_{p=1}^{r} \overline{x_p} \cdot n_p}{n} \\
R &= c_r - c_0 \\
d_x^2 &= \sum_{l=1}^{r} d_l^2 \cdot p_l + \sum_{l=1}^{r} (\overline{x_l} - \overline{x})^2 \cdot p_l = d_a^2 + d_b^2 \\
d_x &= \sqrt{d_x^2} \\
V_x &= \frac{d_x}{\overline{x}}
\end{aligned}
$$

Меры рассеяния (объединённый набор) – Пример из лекции 2

Набор данных: Пример из лекции 2

j	c_u	c_o	n_j	p_j	arithmetisches Mittel	empirische Varianz
1	0	2	58	0.58	0.92	0.3508
2	2	4	22	0.22	2.9	0.3694
3	4	6	16	0.16	4.63	0.2578
4	6	10	3	0.03	7.98	0.038
5	10	14	1	0.01	13.13	0
			100	1		

Примечание:

Средние значения и эмпирические дисперсии вычислены из первичного списка для соответствующих групп.









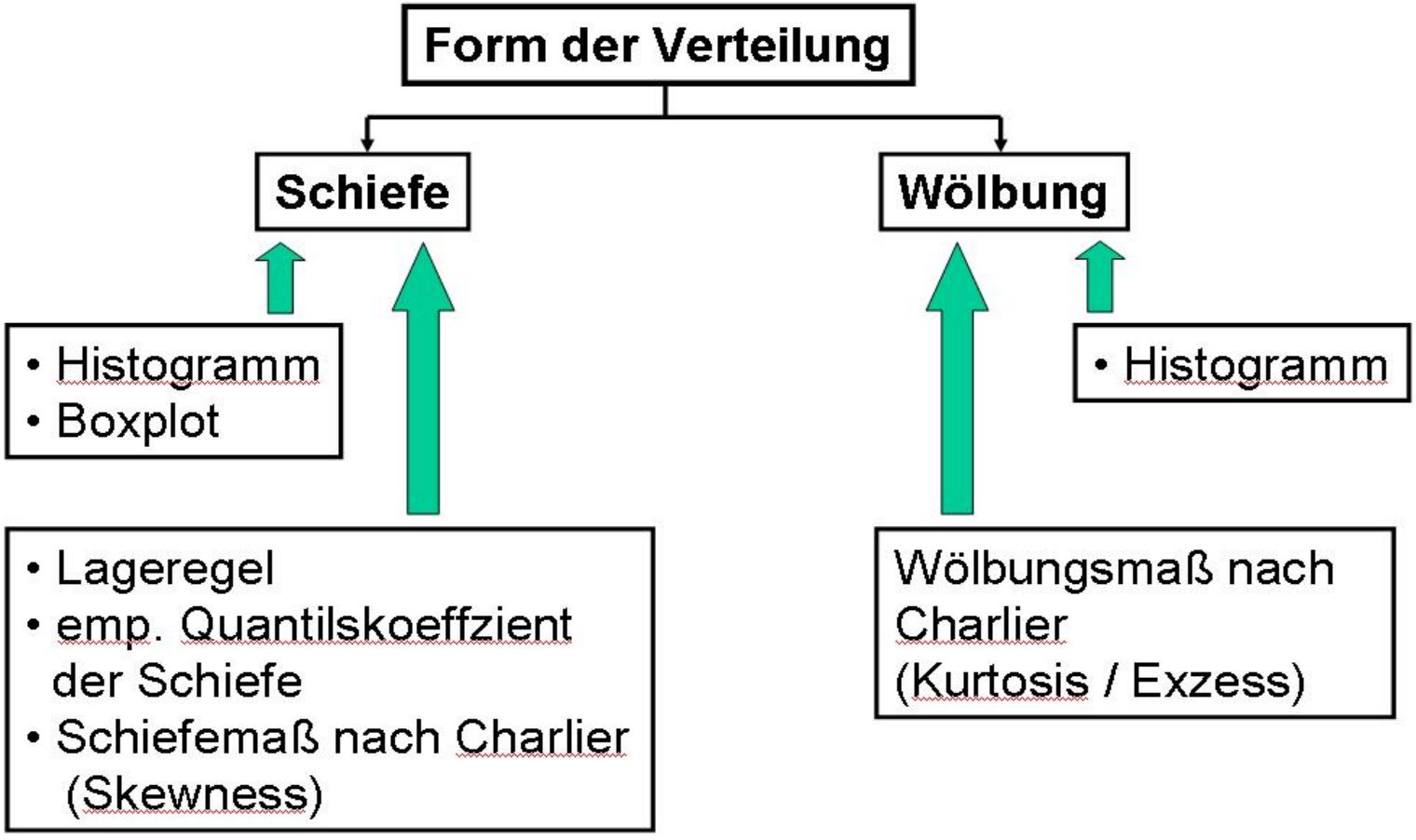
Числовые характеристики распределения

- 2.2.1 Меры положения (меры центральной тенденции)
- 2.2.2 Меры рассеяния
- **2.2.3 Меры формы**
- 2.2.4 Линейные преобразования

2.2.3 Меры формы – Содержание

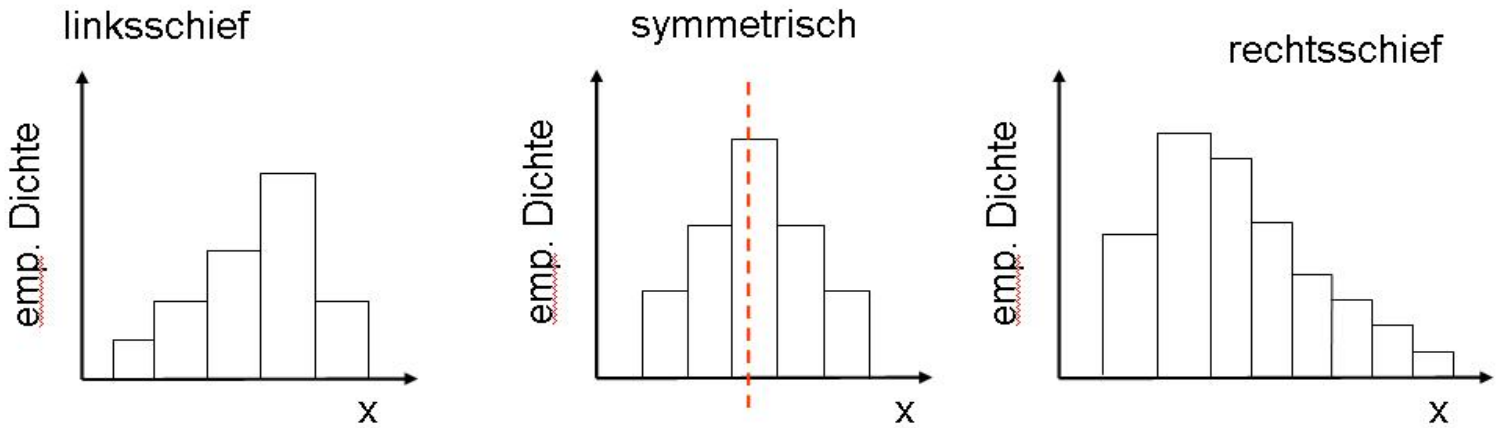
- **2.2.3.1 Асимметрия**
- 2.2.3.2 Эксцесс
- 2.2.3.3 Эмпирические моменты

Меры формы – введение



Симметрия / Асимметрия – гистограмма

Симметрия / асимметрия эмпирического распределения (гистограмма)



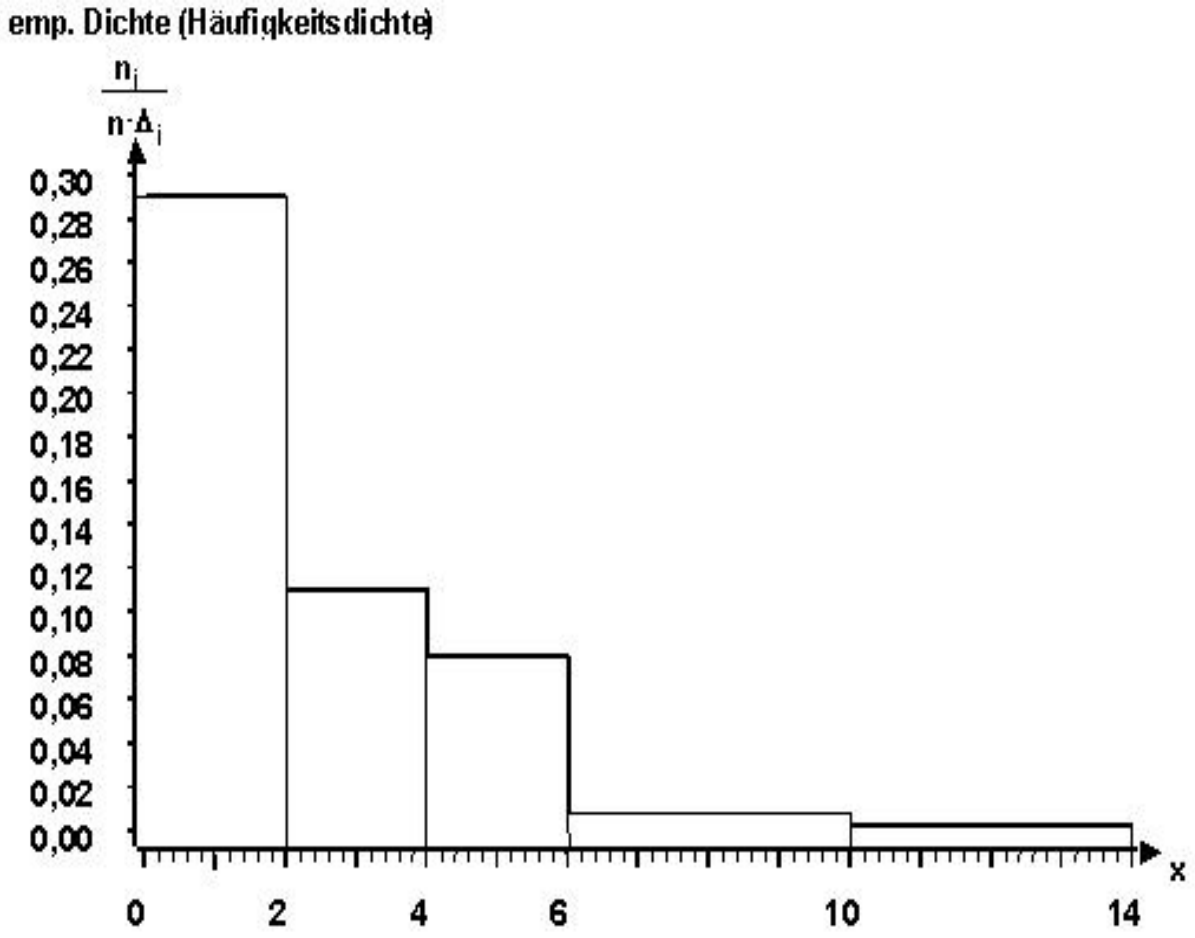
Правило расположения (относительно моды x_M и среднего $\overline{x}$):

$\overline{x} < x_{0,5} < x_M$ $\qquad$ $x_M \cong \overline{x} \cong x_{0,5}$ $\qquad$ $x_M < x_{0,5} < \overline{x}$

Симметрия / Асимметрия – пример

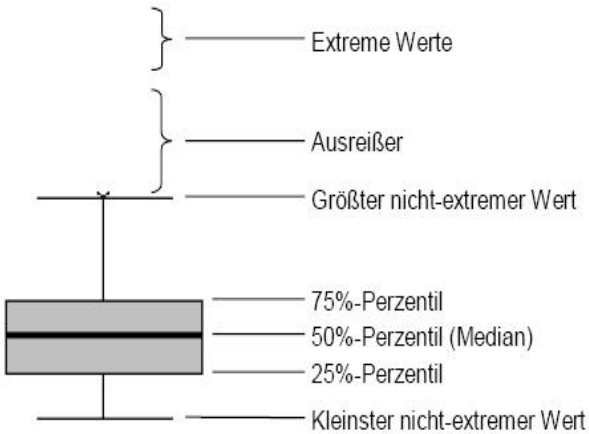
Пример: Телефонные разговоры:

Анализ симметрии/асимметрии эмпирического распределения (гистограмма)



Симметрия / Асимметрия – боксплот

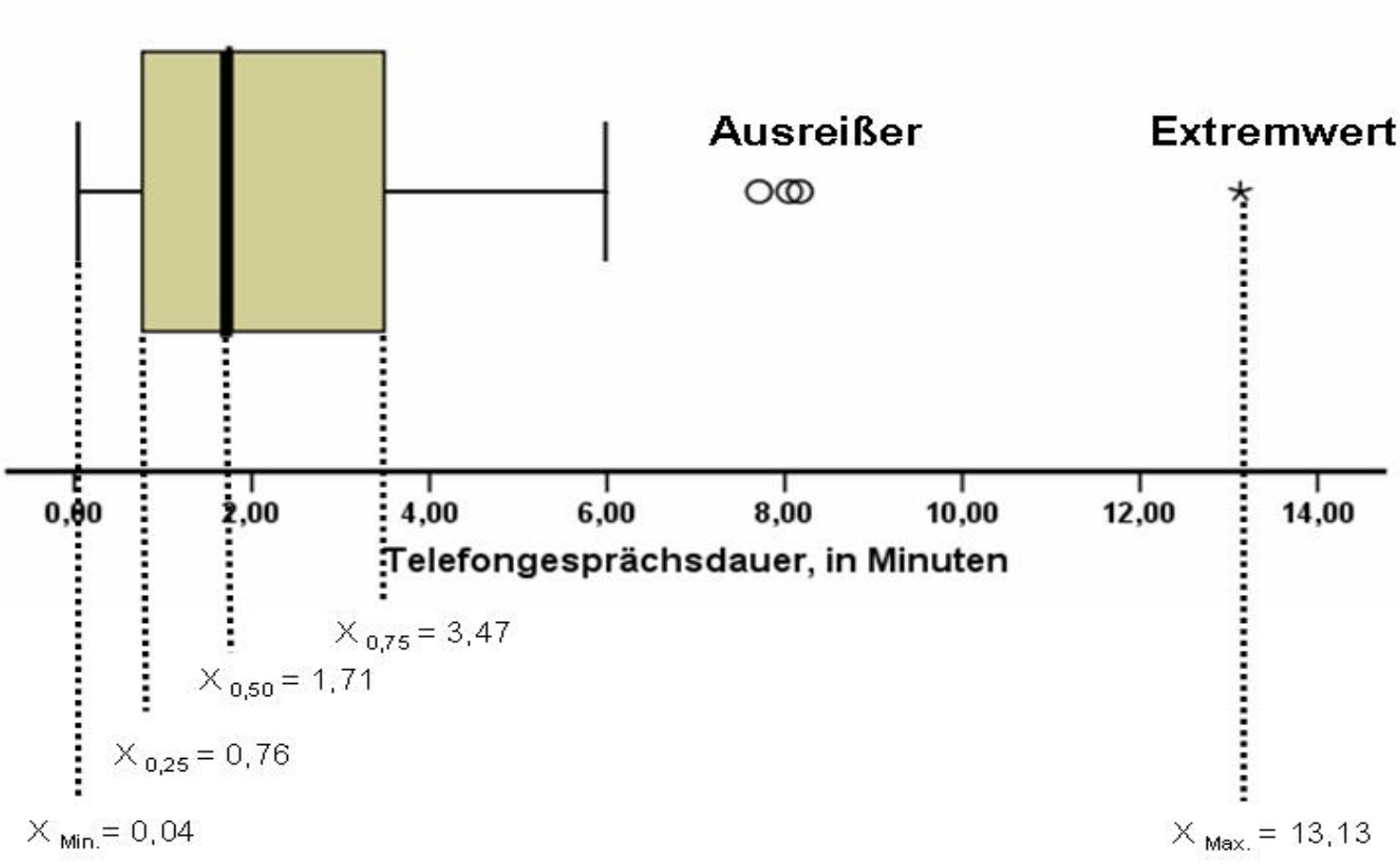
Симметрия / асимметрия эмпирического распределения (Box-and-Whisker-Plot)



Форма и положение ящика могут использоваться для анализа симметрии / асимметрии.

Симметрия / Асимметрия – пример

Пример: Телефонные разговоры:
Анализ симметрии/асимметрии эмпирического распределения (Box-and-Whisker-Plot)



Меры симметрии / асимметрии

Характеристики для измерения асимметрии:
Квантильный коэффициент асимметрии:

$$\frac{(x_{(1+p)/2} - x_{0,5}) - (x_{0,5} - x_{(1-p)/2})}{x_{(1+p)/2} - x_{(1-p)/2}}$$

$p = 0{,}5 \Rightarrow$ Квартильный коэффициент
Квартильный коэффициент асимметрии даёт суждение о 50% средних значений признака:

- если коэффициент < 0 – левосторонняя асимметрия (левый перекос)
- если коэффициент $= 0$ – симметрия
- если коэффициент > 0 – правосторонняя асимметрия (правый перекос)

Симметрия / Асимметрия – Пример из лекции 1

	Несовп. знач. признака	Абс. частоты	Отн. частоты	Абс. кум. частоты
j	ξ_j	n_j	p_j	F_j
1	1	1	0.02	1
2	2	4	0.08	5
3	3	11	0.22	16
4	4	23	0.46	39
5	5	8	0.16	47
6	6	2	0.04	49
7	7	1	0.02	50
Сумма		50	1	

Квантильный коэффициент:
Вычислить квантили:

$$
\begin{aligned}
p = 0.5 &\quad\Rightarrow\quad g = 25 \Rightarrow x_p = 4 \\
p = 0.25 &\quad\Rightarrow\quad g = 12.5 \Rightarrow x_p = 3 \\
p = 0.75 &\quad\Rightarrow\quad g = 37.5 \Rightarrow x_p = 4 \\
p = 0.1 &\quad\Rightarrow\quad g = 5 \Rightarrow x_p = 2 \\
p = 0.9 &\quad\Rightarrow\quad g = 45 \Rightarrow x_p = 6
\end{aligned}
$$

Симметрия / Асимметрия – Пример из лекции 1

$$p = 0.5 \quad : \quad \frac{(x_{\mathbf{0.75}} - x_{\mathbf{0,5}}) - (x_{\mathbf{0,5}} - x_{\mathbf{0,25}})}{x_{\mathbf{0.75}} - x_{\mathbf{0.25}}} = \frac{(4-4)-(4-3)}{4-3} = -1$$

$$p = 0.8 \quad : \quad \frac{(x_{\mathbf{0.9}} - x_{\mathbf{0,5}}) - (x_{\mathbf{0,5}} - x_{\mathbf{0,1}})}{x_{\mathbf{0.9}} - x_{\mathbf{0.1}}} = \frac{(6-4)-(4-2)}{6-2} = 0$$

Симметрия / Асимметрия – Пример из лекции 1

$p = 0.5$: Квартильный коэффициент $= -1$

$p = 0.8$: Квантильный коэффициент $= 0$

Интерпретация:

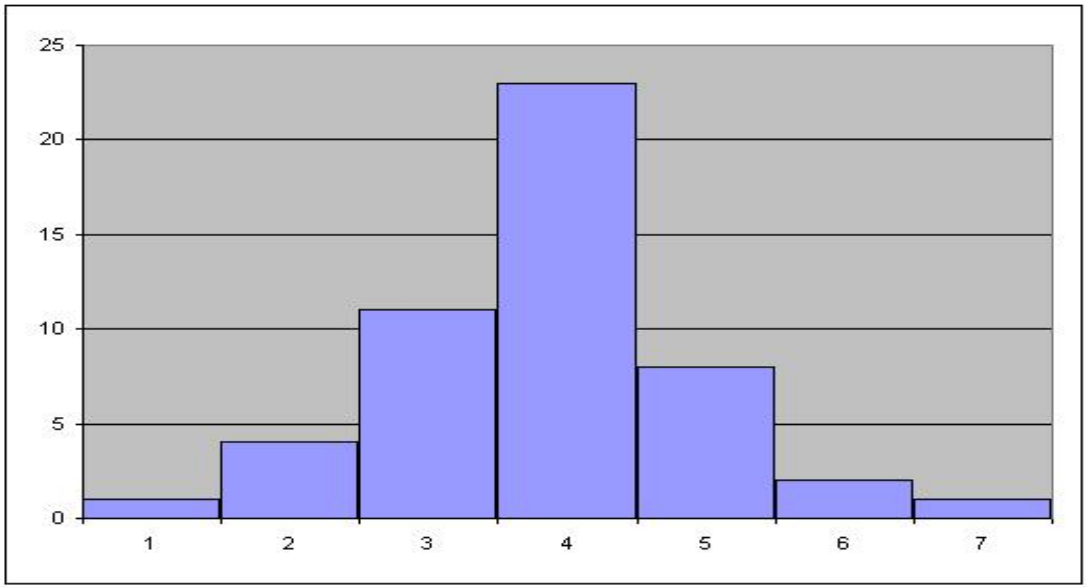
- Квартильный коэффициент $= -1 \Rightarrow$ левосторонняя асимметрия
- Квантильный коэффициент $= 0 \Rightarrow$ симметрия

$\Rightarrow$ противоречивый и грубый результат

см. гистограмму

Симметрия / Асимметрия – Пример из лекции 1

Гистограмма:



Интерпретация: приближённо симметрично, незначительный правый перекос

Меры симметрии / асимметрии

Характеристики для измерения асимметрии (skewness) из первичного списка:
Коэффициент асимметрии по ШАРЛЬЕ S_x:

$$
\begin{aligned}
S_x &= \frac{m_{\overline{x}}^3}{\sqrt{(m_{\overline{x}}^2)^3}} \\
\text{где } m_{\overline{x}}^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \\
m_{\overline{x}}^3 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3
\end{aligned}
$$

Интерпретация:

- если $S_X < 0$ – левосторонняя асимметрия
- если $S_X = 0$ – симметрия
- если $S_X > 0$ – правосторонняя асимметрия

Меры симметрии / асимметрии

Характеристики для измерения асимметрии (skewness) из таблицы частот (несовпадающие значения):
Коэффициент асимметрии по ШАРЛЬЕ S_x:

$$S_x = \frac{m_{\overline{x}}^3}{\sqrt{(m_{\overline{x}}^2)^3}}$$

$$\text{где } m_{\overline{x}}^2 = \sum_{i=1}^{n}(x_i - \overline{x})^2 \cdot p_i$$

$$m_{\overline{x}}^3 = \sum_{i=1}^{n}(x_i - \overline{x})^3 \cdot p_i$$

Интерпретация:

- если $S_X < 0$ – левосторонняя асимметрия
- если $S_X = 0$ – симметрия
- если $S_X > 0$ – правосторонняя асимметрия

Симметрия / Асимметрия – пример

Коэффициент асимметрии по ШАРЛЬЕ S_x из первичного списка:

$$S_x = \frac{m_{\overline{x}}^3}{\sqrt{(m_{\overline{x}}^2)^3}} = \frac{0{,}126912}{\sqrt{1{,}2004^3}} = 0{,}096497$$

$$\text{где } m_{\overline{x}}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = 1{,}2004$$

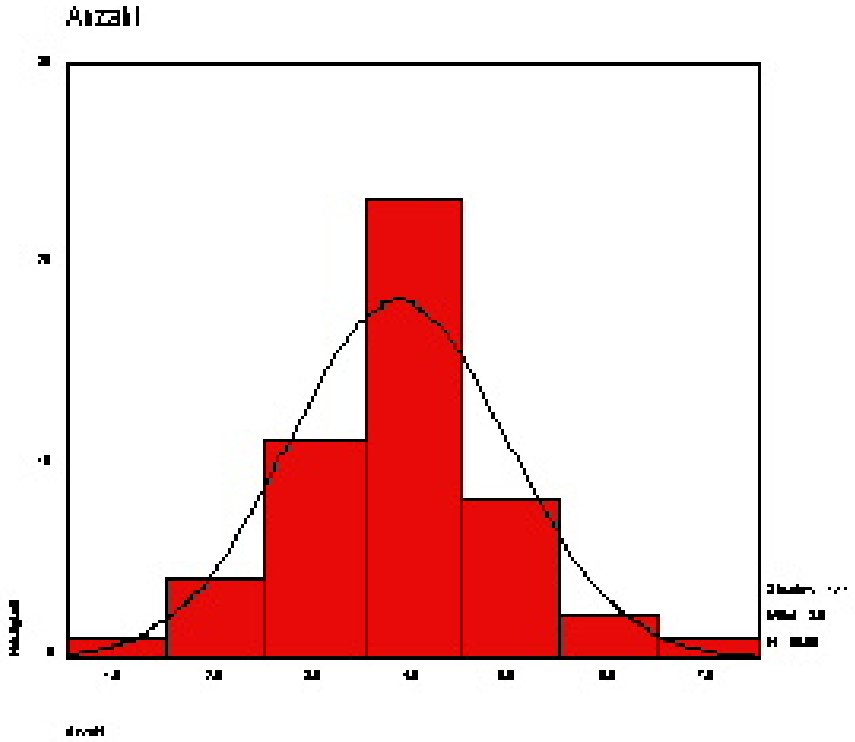
$$m_{\overline{x}}^3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3 = 0{,}126912$$

$S_X \approx 0$ – приближённо симметрично

$S_X > 0$ – незначительный правый перекос

Симметрия / Асимметрия – Пример из лекции 1

Гистограмма с кривой нормального распределения:

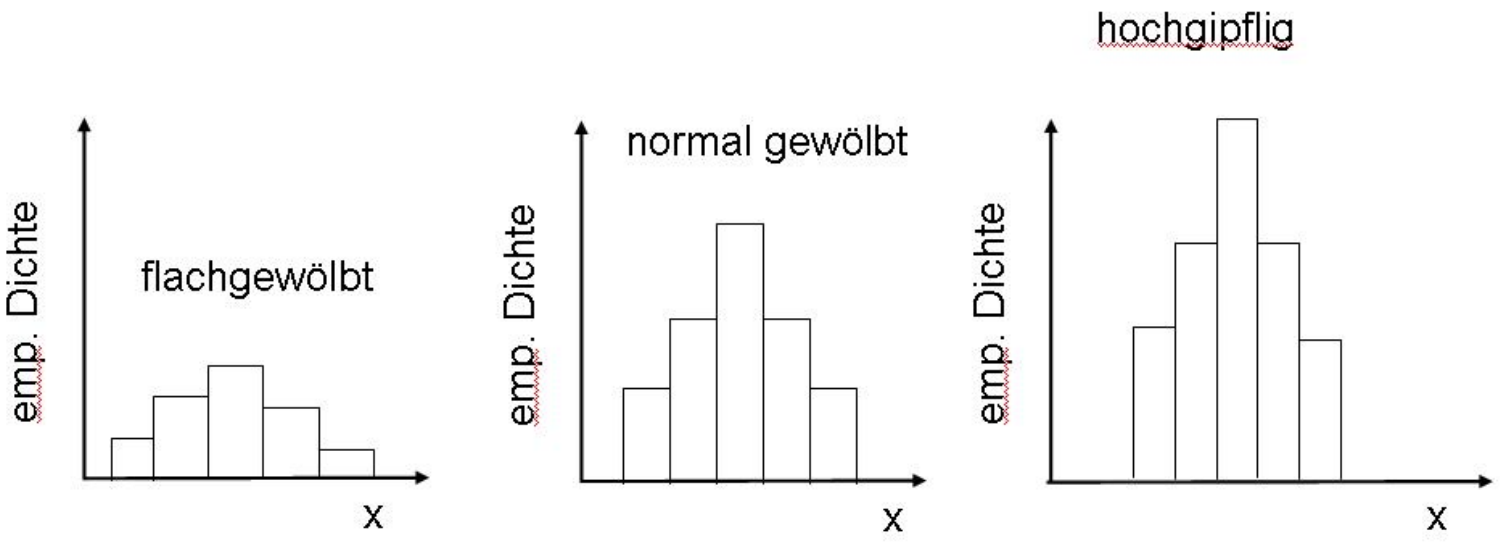


2.2.3 Меры формы – Содержание

- 2.2.3.1 Асимметрия
- **2.2.3.2 Эксцесс**
- 2.2.3.3 Эмпирические моменты

Эксцесс – графическое представление

Эксцесс эмпирического распределения (гистограмма)



Эксцесс – характеристики

Характеристики для измерения эксцесса (kurtosis) из первичного списка:
Мера эксцесса по ШАРЛЬЕ W_x:

$$
\begin{aligned}
W_x &= \frac{m_{\overline{x}}^4}{(m_{\overline{x}}^2)^2} - 3 \\
\text{где } m_{\overline{x}}^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \\
m_{\overline{x}}^4 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^4
\end{aligned}
$$

Интерпретация:

- если $W_X < 0$ – плосковершинное распределение
- если $W_X = 0$ – нормальная вершинность
- если $W_X > 0$ – островершинное распределение

Эксцесс – характеристики

Характеристики для измерения эксцесса (kurtosis) из таблицы частот (несовпадающие значения):
Мера эксцесса по ШАРЛЬЕ W_x:

$$
\begin{aligned}
W_x &= \frac{m_{\overline{x}}^4}{(m_{\overline{x}}^2)^2} - 3 \\
\text{где } m_{\overline{x}}^2 &= \sum_{i=1}^{n} (x_i - \overline{x})^2 \cdot p_i \\
m_{\overline{x}}^4 &= \sum_{i=1}^{n} (x_i - \overline{x})^4 \cdot p_i
\end{aligned}
$$

Интерпретация:

- если $W_X < 0$ – плосковершинное
- если $W_X = 0$ – нормально вершинное
- если $W_X > 0$ – островершинное

Эксцесс – Пример из лекции 1

Мера эксцесса по ШАРЛЬЕ W_x из первичного списка:

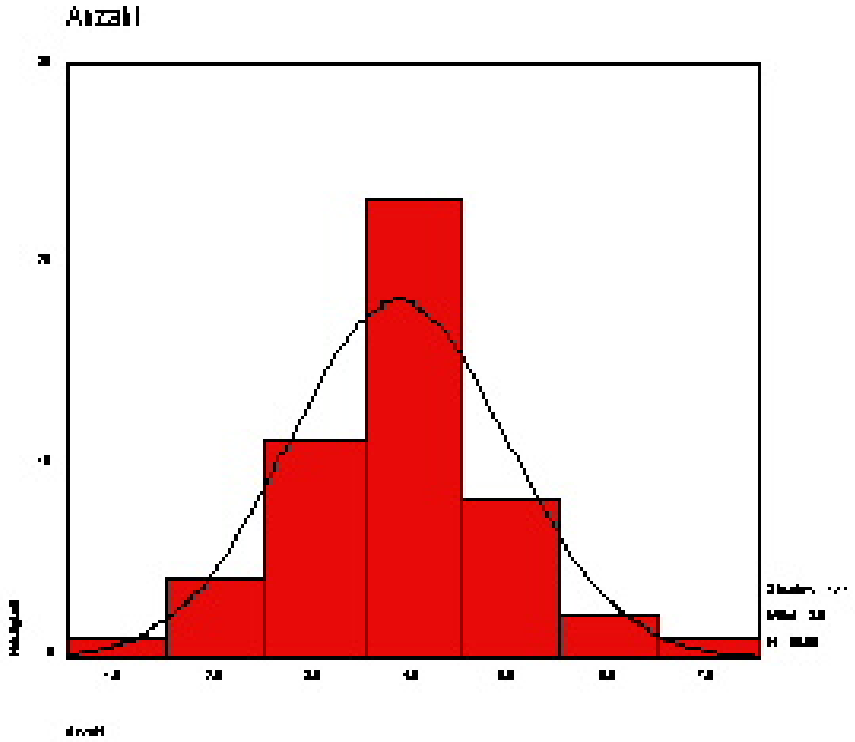
$$
\begin{aligned}
W_x &= \frac{m_{\bar{x}}^4}{(m_{\bar{x}}^2)^2} - 3 = \frac{5{,}46951952}{1{,}2004^2} - 3 = 0{,}795747 \\
\text{где } m_{\bar{x}}^2 &= 1{,}2004 \\
m_{\bar{x}}^4 &= 5{,}46951952
\end{aligned}
$$

Интерпретация:

$W_x = 0{,}795747 > 0 \Rightarrow$ более островершинное, чем нормальное

Эксцесс – Пример из лекции 1

Гистограмма с кривой нормального распределения:



2.2.3 Меры формы – Содержание

- 2.2.3.1 Асимметрия
- 2.2.3.2 Эксцесс
- **2.2.3.3 Эмпирические моменты**

Эмпирические моменты

Эмпирический момент относительно с порядка r:

1. Из первичного списка:

$$m_c^r = \frac{1}{n}\sum_{i=1}^{n}(x_i - c)^r$$

2. Из таблицы частот (без группировки по интервалам):

$$m_c^r = \sum_{l=1}^{m}(\xi_l - c)^r \cdot p_l$$

3. Из таблицы частот (с группировкой по интервалам):

$$m_c^r \approx \sum_{l=1}^{m}(\xi_l^* - c)^r \cdot p_l \text{ где } \xi_l^* \text{ — середина класса}$$

Эмпирические нулевые моменты

Эмпирический момент относительно 0 порядка r, из первичного списка:

$$m_0^r = \frac{1}{n}\sum_{i=1}^{n} x_i^r$$

$r = 1$: первый нулевой момент = среднее арифметическое:

$$m_0^1 = \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Эмпирические центральные моменты

Эмпирический момент относительно $\overline{x}$ порядка r, из первичного списка:

$$m_{\overline{x}}^{r} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^r$$

$r = 2$: второй центральный момент = эмпирическая дисперсия:

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

Другие важные центральные моменты для вычисления асимметрии и эксцесса по ШАРЛЬЕ:

$$r = 3: \quad m_{\overline{x}}^{3} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3$$

$$r = 4: \quad m_{\overline{x}}^{4} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^4$$

Нулевые и центральные моменты – Пример из лекции 1

Первичный список / таблица частот:

$$
\begin{aligned}
m_0^1 &= \overline{x} = 3{,}86 \\
m_{\overline{x}}^2 &= d_X = 1{,}2004 \\
m_{\overline{x}}^3 &= 0{,}126912 \\
m_{\overline{x}}^4 &= 5{,}46951952
\end{aligned}
$$

Нулевые и центральные моменты – Пример из лекции 2

Набор данных с группировкой по интервалам:

$$
\begin{aligned}
m_0^1 &= \overline{x} = \sum_{i=1}^{n} \xi_i^* \cdot p_i = 2{,}4 \\
m_{\overline{x}}^2 &= d_X = \sum_{i=1}^{n} (\xi_i^* - \overline{x})^2 \cdot p_i = 4{,}16 \\
m_{\overline{x}}^3 &= \sum_{i=1}^{n} (\xi_i^* - \overline{x})^3 \cdot p_i = 15{,}384 \\
m_{\overline{x}}^4 &= \sum_{i=1}^{n} (\xi_i^* - \overline{x})^4 \cdot p_i = 124{,}0064 \\
\text{где } \xi_i^* &- \text{середина класса}
\end{aligned}
$$

Введение в SPSS – темы

Статистические темы: дескриптивная статистика

- Частоты – продолжение
- Дескриптивная статистика

Меры формы – пример

с Excel:

Пример 1 (из первичного списка): BeispielW95.xls
Пример 2 (из классифицированных данных): Zinssaetze 2007_1.xls
Пример 3 (объединённый набор): Zinssaetze 2007_4.xls

Задание: определить асимметрию, эксцесс, моменты

с SPSS:

Пример 1 (с группировкой в SPSS): World95.sav.xls
Пример 2 (из классифицированных данных): Zinssaetze 2007_1.xls

Задание: определить асимметрию, эксцесс, моменты

Числовые характеристики распределения

- 2.2.1 Меры положения (меры центральной тенденции)
- 2.2.2 Меры рассеяния
- 2.2.3 Меры формы
- **2.2.4 Линейные преобразования**

2.2.4 Линейные преобразования – Содержание

- **2.2.4.1 Введение**
- 2.2.4.2 Стандартизация
- 2.2.4.3 Стандартизованные характеристики

Линейное преобразование

- x_i – наблюдаемые значения признака
- y_i – преобразованные значения признака
- $y_i = a + b \cdot x_i$ при $b \neq 0$ – правило преобразования

Преобразованные параметры распределения

среднее арифметическое	–	$\overline{y} = a + b\overline{x}$
эмпирическая дисперсия	–	$d_y^2 = b^2 \cdot d_x^2$
эмпирическое стандартное отклонение	–	$d_y = \lvert b \rvert \cdot d_x$
асимметрия	–	$S_y = S_x$
эксцесс	–	$W_y = W_x$

Преобразованные параметры – вывод

$y_i = a + b \cdot x_i$, $b \neq 0$ – правило преобразования

$$\begin{aligned} \overline{y} &= \frac{1}{n}\sum_{i=1}^{n} y_i = a + b \cdot \overline{x} \\ d_y^2 &= \frac{1}{n}\sum_{i=1}^{n} (y_i - \overline{y})^2 = b^2 \cdot d_x^2 \\ d_y &= \sqrt{d_y^2} = |b| \cdot d_x \end{aligned}$$

Преобразованные параметры – вывод

$y_i = a + b \cdot x_i$, $b \neq 0$ – правило преобразования
Асимметрия – $S_y = S_x$

$$
\begin{aligned}
m_{\overline{y}}^3 &= b^3 \cdot m_{\overline{x}}^3 \\
\sqrt{(m_{\overline{y}}^2)^3} &= b^3 \cdot \sqrt{(m_{\overline{x}}^2)^3} \\
S_y &= \frac{m_{\overline{y}}^3}{\sqrt{(m_{\overline{y}}^2)^3}} = \frac{m_{\overline{x}}^3}{\sqrt{(m_{\overline{x}}^2)^3}} = S_x
\end{aligned}
$$

Преобразованные параметры – вывод

$y_i = a + b \cdot x_i$, $b \neq 0$ – правило преобразования
Эксцесс – $W_y = W_x$

$$
\begin{aligned}
m_{\overline{y}}^4 &= b^4 \cdot m_{\overline{x}}^4 \\
(m_{\overline{y}}^2)^2 &= b^4 \cdot (m_{\overline{x}}^2)^2 \\
W_y &= \frac{m_{\overline{y}}^4}{(m_{\overline{y}}^2)^2} - 3 = W_x
\end{aligned}
$$

2.2.4 Линейные преобразования – Содержание

- 2.2.4.1 Введение
- **2.2.4.2 Стандартизация**
- 2.2.4.3 Стандартизованные характеристики

Стандартизация

- x_i – наблюдаемые значения признака
- z_i – преобразованные значения признака
- $z_i = \frac{x_i - \overline{x}}{d_x}$ – правило преобразования

так как

$a = -\frac{\overline{x}}{d_x}$ и $b = \frac{1}{d_x}$

$z_i = \frac{x_i - \overline{x}}{d_x}$ – специальное линейное преобразование

2.2.4 Линейные преобразования – Содержание

- 2.2.4.1 Введение
- 2.2.4.2 Стандартизация
- **2.2.4.3 Стандартизованные характеристики**

Стандартизованные характеристики

$z_i = \frac{x_i - \overline{x}}{d_x}$ – правило преобразования

- $\overline{z} = 0$ – среднее арифметическое
- $d_z = 1$ – стандартное отклонение
- Безразмерные

Стандартизованные характеристики – вывод

$z_i = \frac{x_i - \overline{x}}{d_x}$ – правило преобразования

$$\begin{aligned}
\overline{z} &= \frac{1}{n} \cdot \sum_{i=1}^{n} z_i = \frac{1}{n} \cdot \frac{0}{d_x} = 0 \\
d_z^2 &= \frac{1}{n} \cdot \sum_{i=1}^{n} (z_i - \overline{z})^2 = \frac{1}{d_x^2} \cdot \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \overline{x})^2 = 1 \\
d_z &= \sqrt{d_z^2} = 1
\end{aligned}$$

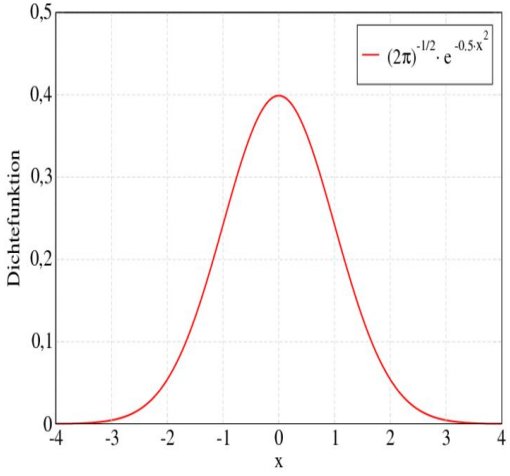
Интерпретация асимметрии – эксцесса

$$\frac{1}{n}\cdot\sum_{i=1}^{n}\left(\frac{x_i-\overline{x}}{d_x}\right)^3 = \frac{1}{n}\cdot\frac{1}{d_x^3}\cdot\sum_{i=1}^{n}(x_i-\overline{x})^3 = \frac{m_{\overline{x}}^3}{\sqrt{(m_{\overline{x}}^2)^3}} = S_x$$

$$\frac{1}{n}\cdot\sum_{i=1}^{n}\left(\frac{x_i-\overline{x}}{d_x}\right)^4 = \frac{1}{n}\cdot\frac{1}{d_x^4}\cdot\sum_{i=1}^{n}(x_i-\overline{x})^4 = \frac{m_{\overline{x}}^4}{(m_{\overline{x}}^2)^2} = W_x+3$$

Экскурс: Нормальное распределение и характеристики распределения

Нормальное распределение ≠ идеально симметричное распределение с нормальным эксцессом
Стандартное нормальное распределение: дополнительно (среднее = 0, стандартное отклонение = 1)



Параметры распределения:

Среднее = 0
Стандартное отклонение = 1
Асимметрия = 0
Эксцесс = 0