Статистика для специалистов по бизнес-информатике
Одномерная статистика – Распределение частот

Д-р Вольфганг Бёме

Октябрь 2010

2. Одномерный анализ распределений – Содержание

- 2.0 Введение
- **2.1 Распределение частот / Функция распределения**
- 2.2 Числовые характеристики распределения
- 2.3 Анализ концентрации

2.1 Распределение частот / Функция распределения – Содержание

- **2.1.1 Распределения частот**
- 2.1.2 Функции распределения

Намерение

Сжатие (компрессия) первичного списка путём анализа значений признака

Несовпадающие значения признака ξ_l $(l = 1, 2, ..., m)$
с $\xi_{l_1} \neq \xi_{l_2}$ $\forall l_1, l_2$ *mit* $l_1 \neq l_2$

Носитель признака	Значение признака	Несовпадающее значение признака
1	x_1	ξ_1
2	x_2	ξ_1
3	x_3	ξ_2
4	x_4	ξ_1
5	x_5	ξ_2
6	x_6	ξ_3
$\vdots$	$\vdots$	$\vdots$

2.1.1 Распределения частот – Содержание

- **2.1.1.1 Простые распределения частот**
- 2.1.2.2 Распределения частот с группировкой по интервалам

Абсолютное распределение частот

Упорядочение несовпадающих значений признака: $\xi_{l_1} > \xi_{l_2}$ *wenn* $l_1 > l_2$ и определение абсолютных частот n_l:

$n_l = n(\xi_l) = n(X = \xi_l)$ – число значений признака $X = \xi_l$

$$n = \sum_{l=1}^{m} n_l$$

Определение: Абсолютное распределение частот – это соответствие несовпадающих значений признака их абсолютным частотам:

$$f : \xi_l \to n_l = f(\xi_l) = n(\xi_l)$$

Относительное распределение частот

Исходная точка: абсолютные частоты n_l:

$n_l = n(\xi_l)$ – число значений признака $X = \xi_l$

Относительная частота p_l:

$$\begin{aligned} p_l &= p(X = \xi_l) = \frac{n_l}{n} \\ n &= \sum_{l=\mathbf{1}}^{m} n_l \Rightarrow \\ \mathbf{1} &= \sum_{l=\mathbf{1}}^{m} p_l \end{aligned}$$

Интерпретация: доля значений признака $(X = \xi_l)$ в общем числе всех значений признака n.

Определение: Относительное распределение частот – это соответствие несовпадающих значений признака их относительным частотам:

$$f : \xi_l \to p_l = f(\xi_l) = p(\xi_l)$$

Таблица распределения частот

Несовпадающее значение признака	Абсолютные частоты	Относительные частоты
ξ_1	n_1	p_1
ξ_2	n_2	p_2
$\vdots$	$\vdots$	$\vdots$
ξ_m	n_m	p_m

Пример: Распределение частот

Набор данных: Vorlesungsbeispiel1.xls

	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

Графическое представление распределения частот

- Столбчатая диаграмма
- Линейная диаграмма
- Круговая диаграмма

⇒ Графические представления в Excel
(см. Vorlesungsbeispiel1.xls)

⇒ Графические представления в SPSS

2.1.1 Распределения частот – Содержание

- 2.1.1.1 Простые распределения частот
- **2.1.2.2 Распределения частот с группировкой по интервалам**

Абсолютное распределение частот с группировкой по интервалам

Вместо несовпадающих значений признака можно образовывать несовпадающие классы признака: Область значений признака X первичного списка делится значениями $c_0 < c_1 < ... < c_r$ на r интервалов:

$$\begin{aligned} I_1 &= [c_0, c_1] \\ I_2 &= (c_1, c_2] \\ ... &\quad ... \\ I_r &= (c_{r-1}, c_r] \end{aligned}$$

Определение абсолютных частот n_I:

$n_I = n(I_r)$ – число значений признака $X \in I_r$

Определение: Абсолютное распределение частот – это соответствие несовпадающих классов признака (интервалов) их абсолютным частотам:

$$f : I_r \to n_I = f(I_r) = n(I_r)$$

Относительное распределение частот с группировкой по интервалам

Исходная точка: абсолютные частоты n_l:

$n_l = n(I_r)$ – число значений признака $X \in I_r$

Относительная частота p_l:

$$\begin{aligned} p_l &= P(X \in I_r) = \frac{n_l}{n} \\ n &= \sum_{l=\mathbf{1}}^{m} n_l \Rightarrow \\ \mathbf{1} &= \sum_{l=\mathbf{1}}^{m} p_l \end{aligned}$$

Интерпретация: доля значений признака $(X \in I_r)$ в общем числе всех значений признака n.

Определение: Относительное распределение частот – это соответствие несовпадающих классов признака (интервалов) их относительным частотам:

$$f : I_r \to p_l = f(I_r) = p(I_r)$$

Пример из лекции 2

Vorlesungsbeispiel 2.1-b
Statistische Masse: die einem Privathaushalt in Rechnung gestellten Telefonverbindungen
Erhebungsmerkmal: Dauer der Verbindung, in Minuten

Urliste :

1,72	2,82	3,90	4,23	1,26
0,51	7,71	1,42	1,57	0,10
0,66	1,33	026	1,06	0,39
4,14	3,51	0,76	4,60	5,42
2,92	13,13	1,99	8,17	0,77
0,14	3,23	3,86	2,59	0,27
0,97	4,28	0,87	2,15	2,14
1,57	4,32	1,20	0,38	1,19
0,10	1,33	4,32	0,04	2,02
4,41	1,29	3,11	1,62	1,74
0,09	4,68	3,64	2,58	1,05
4,67	0,97	1,07	0,51	3,15
1,68	2,13	3,60	8,05	0,23
1,93	4,41	0,12	2,01	4,59
0,27	5,39	4,11	5,98	2,28
3,11	4,52	0,55	1,78	2,38
1,79	0,48	0,16	0,97	3,47
0,87	0,22	3,13	0,34	1,70
1,54	1,15	0,46	0,63	1,74
0,11	1,03	1,86	0,44	0,84

Пример из лекции 2

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

Пример из лекции 2

Набор данных: Vorlesungsbeispiel2.xls (Телефонное соединение; число классов $m = 5$)

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

2.1 Распределение частот / Функция распределения – Содержание

- 2.1.1 Распределения частот
- **2.1.2 Функции распределения**

2.1.2 Функции распределения – Содержание

- **2.1.1.1 Простые функции распределения**
- 2.1.2.2 Функции распределения с группировкой по интервалам
- 2.1.2.3 Гистограммы

Абсолютное кумулятивное распределение частот

Исходная точка: абсолютные частоты n_l:

$n_l = n(\xi_l)$ – число значений признака $X = \xi_l$

Абсолютная кумулятивная частота H_l:

$$H_l \quad = \quad H(X \leq \xi_l) = \sum_{r=1}^{l} n_r$$

Определение: **Абсолютное кумулятивное распределение частот** – это соответствие несовпадающих значений признака абсолютным кумулятивным частотам:

$$f : \xi_l \to H_l = f(\xi_l) = H(X \leq \xi_l)$$

Примечание: признак X измерен хотя бы по порядковой шкале, первичный список отсортирован по возрастанию.

Относительное кумулятивное распределение частот

Исходная точка: относительные частоты p_l:

$p_l = p(\xi_l)$ – относительная частота значений признака $X = \xi_l$
Относительная кумулятивная частота F_l:

$$F_l \quad = \quad F(X \leq \xi_l) = \frac{H(X \leq \xi_l)}{n} = \sum_{r=1}^{l} p_r$$

Определение: **Относительное кумулятивное распределение частот** – это соответствие несовпадающих значений признака относительным кумулятивным частотам:

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

Примечание: признак X измерен хотя бы по порядковой шкале, первичный список отсортирован по возрастанию.

Функция распределения

Исходная точка: относительные частоты p_l: $p_l = p(\xi_l)$ – относительная частота значений признака $X = \xi_l$

$$F(x) = \begin{cases} 0 & \forall x : x < \xi_1 \\ F_l = \sum\limits_{r=1}^{l} p_r & \forall x : \xi_l \le x < \xi_{l+1} \quad l = 1, 2, ..., m-1 \\ 1 & \forall x : x \ge \xi_m \end{cases}$$

где

x – вещественное числоξ_1 - наименьшее наблюдаемое значение

ξ_m – наибольшее наблюдаемое значение

Пример из лекции 1

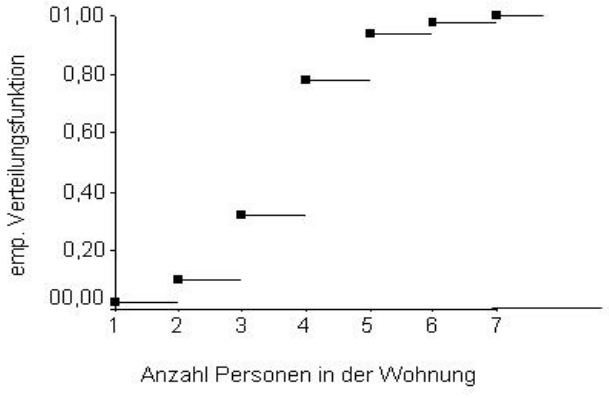
Urliste **aufsteigend sortierte Urliste**

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

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

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

Интерпретация функций распределения

- $P(a \leq X \leq b)$ – относительная частота всех значений признака X, для которых: $a \leq X \leq b$.
- $F(x)$ – значение функции распределения в точке x.

  Содержательно: относительная частота всех значений признака X, для которых: $-\infty < X \leq x$, т.е. относительная кумулятивная частота

  $P(X \leq x) = F(x)$
- $P(X < x) = P(X \leq x) - p(X = x) = F(x) - p(X = x)$
  $p(X = x)$ – относительная частота того, что $X = x$
- $P(X > x) = 1 - P(X \leq x) = 1 - F(x)$
- $P(a < X \leq b) = F(b) - F(a)$
- $P(a \leq X \leq b) = F(b) - F(a) + p(X = a)$
  $p(X = x)$ – относительная частота того, что $X = x$
- $P(a < X < b) = F(b) - F(a) - p(X = b)$

Некоторые вопросы и интерпретации

$$
\begin{aligned}
P(a \leq X \leq b) &= F(b) - F(a) + p(X = a) \\
P(3 \leq X \leq 5) &= F(5) - F(3) + p(X = 3) \\
P(3 \leq X \leq 5) &= 0{,}94 - 0{,}32 + 0{,}22 = \\
P(3 \leq X \leq 5) &= 0{,}84 \\
P(3 \leq X \leq 5) &= p(X = 3) + p(X = 4) + p(X = 5) = \\
P(3 \leq X \leq 5) &= 0{,}22 + 0{,}46 + 0{,}16 = \\
P(3 \leq X \leq 5) &= 0{,}84
\end{aligned}
$$

Некоторые вопросы и интерпретации

1. $F(2) = ?$
2. $F(2{,}5) = ?$
3. $F(4{,}5) = ?$
4. $F(4{,}5) - F(2{,}5) = ?$
5. $F(10) = ?$
6. $F(0) = ?$

2.1.2 Функции распределения – Содержание

- 2.1.1.1 Простые функции распределения
- **2.1.2.2 Функции распределения с группировкой по интервалам**
- 2.1.2.3 Гистограммы

Абсолютное кумулятивное распределение – с группировкой по интервалам

Исходная точка: абсолютные частоты n_l:

$n_l = n(I_l)$ – число значений признака $X \in I_l$ с $I_l = (c_{l-\mathbf{1}}, c_l]$

Абсолютная кумулятивная частота H_l:

$$H_l \quad = \quad H(X \leq c_l) = \sum_{r=\mathbf{1}}^{l} n_r$$

Определение: Абсолютное кумулятивное распределение частот – это соответствие верхних границ несовпадающих классов признака (интервалов) абсолютным кумулятивным частотам:

$$f : c_l \to H_l = f(c_l) = H(X \leq c_l)$$

Примечание: признак X измерен хотя бы по порядковой шкале, первичный список отсортирован по возрастанию.

Относительное кумулятивное распределение – с группировкой по интервалам

Исходная точка: относительные частоты p_l:

$p_l = p(I_l)$ – относительная частота значений признака $X \in I_l$ с $I_l = (c_{l-\mathbf{1}}, c_l]$
Относительная кумулятивная частота F_l:

$$F_l \quad = \quad F(X \leq c_l) = \frac{H(X \leq c_l)}{n} = \sum_{r=\mathbf{1}}^{l} p_r$$

Определение: Относительное кумулятивное распределение частот – это соответствие верхних границ несовпадающих классов признака (интервалов) относительным кумулятивным частотам:

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

Примечание: признак X измерен хотя бы по порядковой шкале, первичный список отсортирован по возрастанию.

Эмпирическая функция распределения – с группировкой по интервалам

Исходная точка: относительные частоты p_l:

X – классифицированный признак, измеренный по кардинальной шкале

$p_l = p(I_r)$ – относительная частота значений признака $X \in I_l$

$$F(x) = \begin{cases} 0 & \forall x : x \leq c_{\mathbf{0}} \\ F_{l-\mathbf{1}} + p_l^D (x - c_{l-\mathbf{1}}) & \forall x : c_{l-\mathbf{1}} < x \leq c_l \\ 1 & \forall x : x > c_m \end{cases} \qquad l = 1, 2, \ldots, m$$

где

$$\begin{aligned} x &\quad - \quad \text{вещественное число} \\ F_{l-\mathbf{1}} &= F(X \leq c_{l-\mathbf{1}}) = \sum_{r=\mathbf{1}}^{l-\mathbf{1}} p_r \\ \Delta_l &= c_l - c_{l-\mathbf{1}} \\ p_l^D &= \frac{p_l}{c_l - c_{l-\mathbf{1}}} = \frac{p_l}{\Delta_l} \\ \Delta_l &\quad - \quad \text{ширина интервала} \\ p_l^D &\quad - \quad \text{эмпирическая плотность (плотность частот)} \end{aligned}$$

Эмпирическая функция распределения – интерпретация

- Эмпирическая плотность p_l^D

  $p_l^D = \frac{p_l}{c_l - c_{l-1}} = \frac{p_l}{\Delta_l}$

- Равномерное распределение в классе:

  $F_{l-1} + p_l^D(x - c_{l-1}) = F_{l-1} + \frac{p_l}{c_l - c_{l-1}}(x - c_{l-1})$

Пример из лекции 2

Аналитическое представление эмпирической функции распределения (с группировкой по интервалам)

Пример из лекции 2 (Телефонное соединение; число классов $m = 5$)

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. -obergrenze

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

Пример из лекции 2

Аналитическое представление эмпирической функции распределения (с группировкой по интервалам)

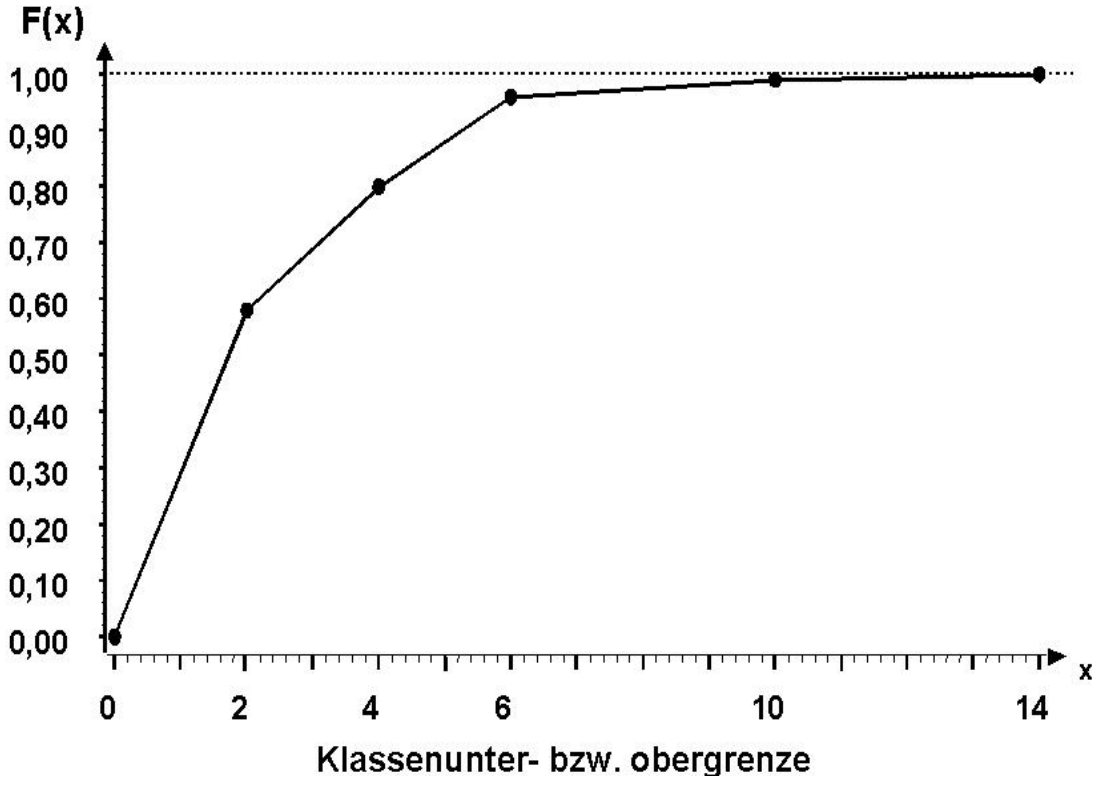
Пример из лекции 2 (Телефонное соединение; число классов $m = 5$)

$$F(x) = \begin{cases} 0 & \text{falls} \quad x \le 0 \\ 0{,}29 \cdot x & \text{falls} \quad 0 < x \le 2 \\ 0{,}58 + 0{,}11 \cdot (x-2) & \text{falls} \quad 2 < x \le 4 \\ 0{,}80 + 0{,}08 \cdot (x-4) & \text{falls} \quad 4 < x \le 6 \\ 0{,}96 + 0{,}0075 \cdot (x-6) & \text{falls} \quad 6 < x \le 10 \\ 0{,}99 + 0{,}0025 \cdot (x-10) & \text{falls} \quad 10 < x \le 14 \\ 1{,}00 & \text{falls} \quad x > 14 \end{cases}$$

Пример из лекции 2

Графическое представление эмпирической функции распределения (с группировкой по интервалам)

Пример из лекции 2 (Телефонное соединение; число классов $m = 5$)



Интерпретация функций распределения

- $P(a \leq X \leq b)$ – относительная частота всех значений признака X, для которых: $a \leq X \leq b$.
- $F(x)$ – значение функции распределения в точке x.

  Содержательно: относительная частота всех значений признака X, для которых: $-\infty < X \leq x$, т.е. относительная кумулятивная частота

  $P(X \leq x) = F(x)$
- $P(X > x) = 1 - P(X \leq x) = 1 - F(x)$
- $P(a < X \leq b) = P(a \leq X \leq b) = P(a < X < b) = F(b) - F(a)$

Некоторые вопросы и интерпретации

$$
\begin{aligned}
P(a \leq X \leq b) &= F(b) - F(a) \\
P(3 \leq X \leq 5) &= F(5) - F(3) \\
P(3 \leq X \leq 5) &= 0{,}88 - 0{,}69 = \\
P(3 \leq X \leq 5) &= 0{,}19 \\
P(3 \leq X < 5) &= 0{,}19 \\
P(3 < X \leq 5) &= 0{,}19 \\
P(3 < X < 5) &= 0{,}19
\end{aligned}
$$

Некоторые вопросы и интерпретации

1. $F(2) = ?$
2. $F(2{,}5) = ?$
3. $F(4{,}5) = ?$
4. $F(4{,}5) - F(2{,}5) = ?$
5. $F(10) = ?$
6. $F(0) = ?$

2.1.2 Функции распределения – Содержание

- 2.1.1.1 Простые функции распределения
- 2.1.2.2 Функции распределения с группировкой по интервалам
- **2.1.2.3 Гистограммы**

Гистограммы

Гистограмма – это графическое представление распределения частот измеренных значений. При этом исходят из данных, упорядоченных по величине, и делят всю область выборки на k интервалов I_l. Они не обязаны быть одинаковой ширины. Однако одинаковые классы, по крайней мере в средней части, упрощают интерпретацию.
Над каждым классом строится прямоугольник, площадь которого пропорциональна классовой частоте.
Если площадь прямоугольника равна абсолютной частоте, гистограмма называется абсолютной; если используются относительные частоты – относительной или нормированной.

Высота прямоугольника h_l:

$$\begin{aligned} \text{абсолютная} \quad &: \quad h_l = \frac{n_l}{\Delta_l} \\ \text{относительная} \quad &: \quad h_l = \frac{n_l}{n}\frac{1}{\Delta_l} = p_l\frac{1}{\Delta_l} \\ \text{где } \Delta_l \quad &- \quad \text{ширина интервала} \end{aligned}$$

Ширина прямоугольника = ширина интервала

Гистограммы

так как

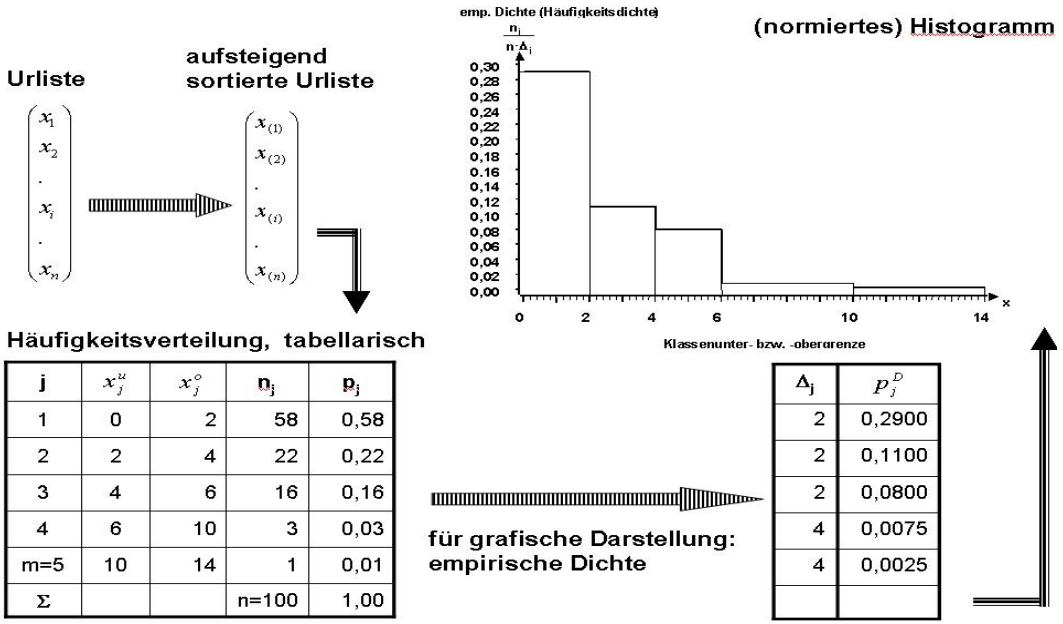
$$
\begin{aligned}
1 &= \sum_{l=1}^{k} p_l \text{ и} \\
h_l &= \frac{n_l}{n}\frac{1}{\Delta_l} = p_l \frac{1}{\Delta_l} \Rightarrow \\
1 &= \sum_{l=1}^{k} h_l \Delta_l
\end{aligned}
$$

Для относительной гистограммы: суммарная площадь всех прямоугольников равна 1.

Пример из лекции 2

Гистограмма эмпирической функции распределения (с группировкой по интервалам)

Пример из лекции 2 (Телефонное соединение; число классов $m = 5$)



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

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