Chapter II

Distribution Analysis

D. 2. 1. (***Absolute and Relative Frequencies***)
Let X be a characteristic possessing the attributes a_j, $j = 1, 2,..., k$.
The *absolute frequency* of the attribute a_j, $j = 1, 2,..., k$ is defined as follows:

$F(a_j) :=$ "Number of cases in which a_j occurs", $j = 1, 2,..., k$.

The *relative frequency* of the attribute a_j, $j = 1, 2,..., k$, is defined as:

$$f(a_j) := \frac{1}{n} \cdot F(a_j), \quad j = 1, 2,..., k.$$

n: number of observations.

R. 2. 1.

$$0 \le F(a_j) \le n, \qquad j = 1,\ 2,..., k$$

$$0 \le f(a_j) \le 1,\ j = 1,\ 2,..., k$$

$$\sum_{j=1}^{k} F(a_j) = n,$$

$$\sum_{j=1}^{k} f(a_j) = 1.$$

Ex. 2. 1.
Denote by

X : wages per hour in € of 30 workers in a firm:

17.05,	17.80,	17.80,	14.70,	15.15,	18.30,
16.20,	16.20,	16.55,	17.05,	15.15,	15.60,
15.60,	15.60,	16.20,	14.00,	15.00,	15.60,
16.55,	18.00,	17.50,	17.05,	16.55,	15.60,
15.60,	15.00,	16.20,	18.30,	17.50,	15.60.

$n = 30$, $\qquad k = 12$.

a_j	$F(a_j)$	$f(a_j)$
14,00	1	0,033
14,70	1	0,033
15,00	2	0,067
15,15	2	0,067
15,60	7	0,233
16,20	4	0,133
16,55	3	0,100
17,05	3	0,100
17,50	2	0,067
17,80	2	0,067
18,00	1	0,033
18,30	2	0,067
Total	30	1.000

D. 2. 2. (***Absolute and Relative Frequency Distributions***)
The sequence

$$F(a_1),\ F(a_2),\ ...,\ F(a_k)$$

is defined as the *absolute frequency distribution*.

The sequence

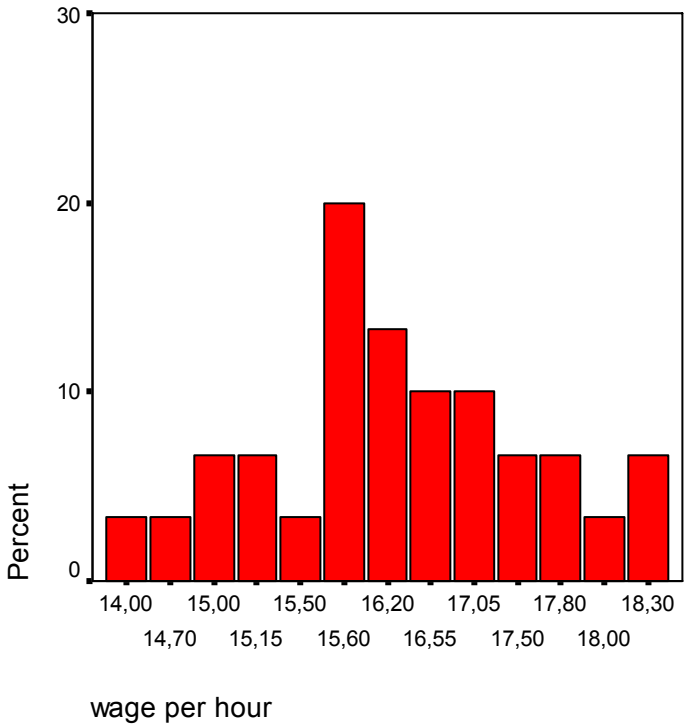
$$f_n(a_1),\ f(a_2),\ ...,\ f(a_k)$$

is defined as the *relative frequency distribution*.

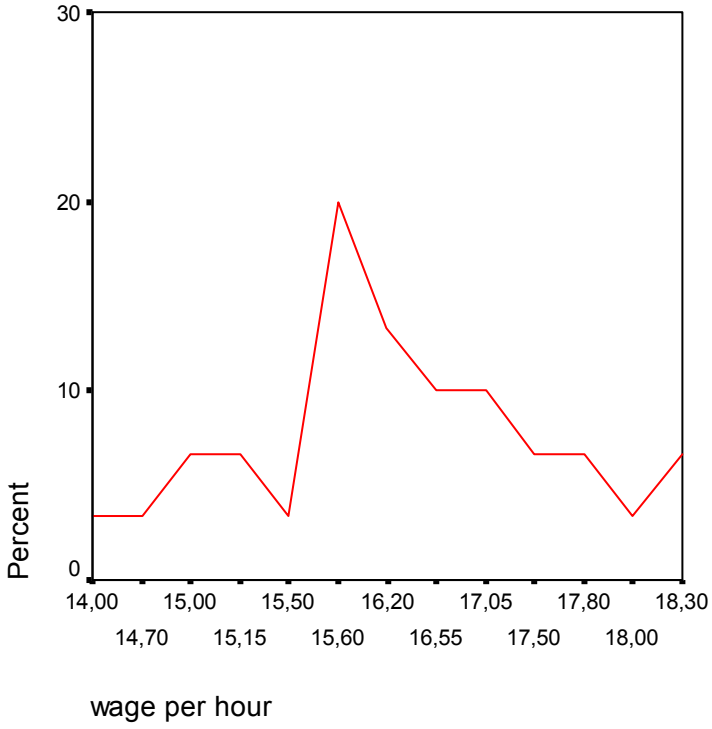
R. 2. 2. (***Graphical Representation of Frequencies***)
We distinguish

1. *bar charts*
   A bar chart depicts the frequencies by a series of bars.
2. *polygons*
   A frequency polygon is a line graph of a frequency distribution.
3. *frequency curves*
   A frequency curve is a smoothed frequency polygon.
4. *histograms*
   A histogram is a bar graph of a frequency distribution
5. *pie charts* (*circle diagrams*)
   A pie chart is a pie-shaped figure in which pieces of the pie represent the frequencies.

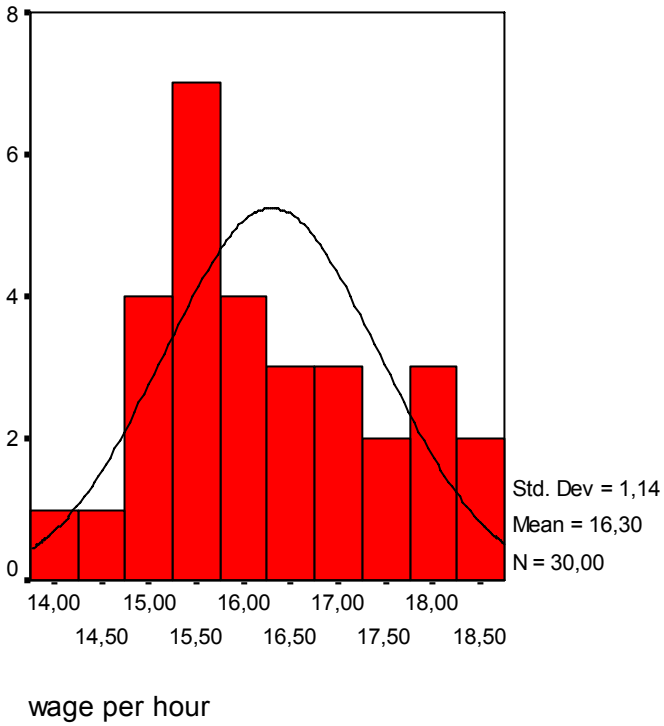
Ex. 2. 1. (*cont.*)
1. Bar chart



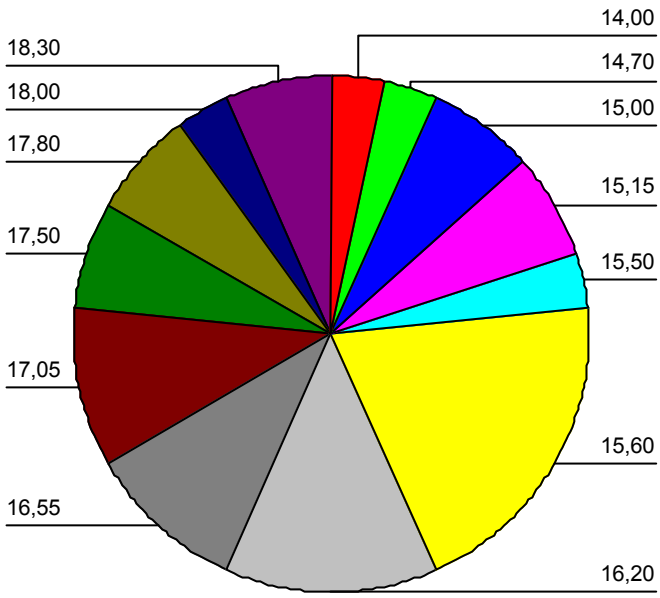
2. Polygon



4. Histogramm



5. Pie chart



R. 2. 3.

In terms of *skewness,* a frequency curve can be

1. *negatively skewed*: nonsymmetrical with the « tail » to the left.
2. *positively skewed*: nonsymmetrical with the « tail » to the right.
3. *symmetrical*.

In terms of *kurtosis*, a frequency curve can be

1. *platykurtic*: flat with the observations distributed relatively evenly.
2. *leptokurtic*: peaked with the observations concentrated within a narrow range of values.
3. *mesokurtic*: neither flat nor peaked, in terms of the distribution of observed values.

D. 2. 3. (***Cumulative Absolute and Relative Frequencies***)
The *cumulative absolute frequency* of the attribute a_j, $j = 1, 2, ..., k$, is defined as:

$$F(a_1) + F(a_2) + ... + F(a_j) = \sum_{i=1}^{j} F(a_i), \qquad j = 1, 2, ..., k .$$

The *cumulative relative frequency* of the attribute a_j, $j = 1, 2, ..., k$, is defined as:

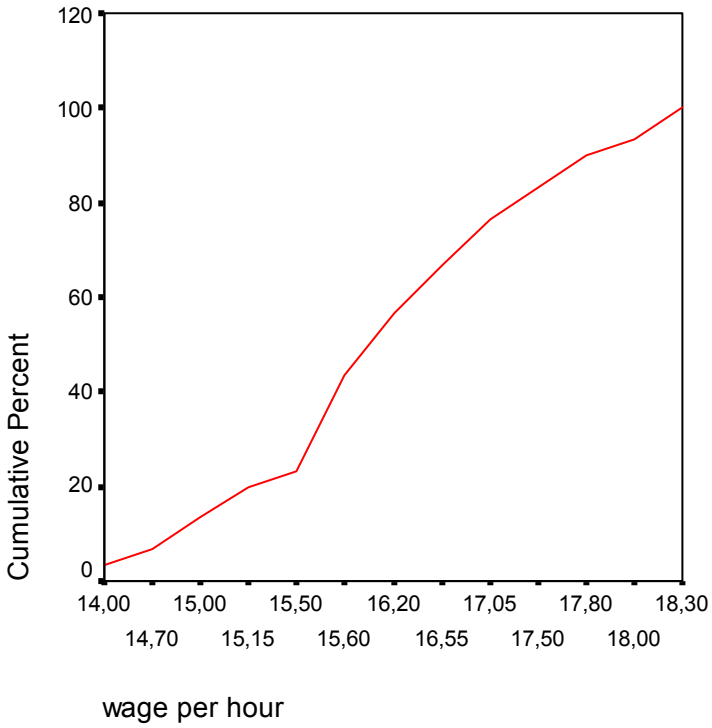
$$f(a_1) + f(a_2) + ... + f(a_j) = \sum_{i=1}^{j} f(a_i), \qquad j = 1, 2, ..., k$$

Ex. 2. 1. (*cont.*)

j	a_j	$F(a_j)$	$f(a_j)$	$\sum_{i=1}^{j} F(a_i)$	$\sum_{i=1}^{j} f(a_i)$
1	14,00	1	0,033	1	0,033
2	14,70	1	0,033	2	0,066
3	15,00	2	0,067	4	0,133
4	15,15	2	0,067	6	0,200
5	15,60	7	0,233	13	0,433
6	16,20	4	0,133	17	0,566
7	16,55	3	0,100	20	0,666
8	17,05	3	0,100	23	0,766
9	17,50	2	0,067	25	0,833
10	17,80	2	0,067	27	0,900
11	18,00	1	0,033	28	0,933
12	18,30	2	0,067	30	1,000
	Total	30	1.000		

R. 2. 4. (***Ogive***)
An *ogive* shows data values on the horizontal axis and either the cumulative absolute frequencies, the cumulative relative frequencies, or the cumulative percent frequencies on the vertical axis.

Ex. 2. 1. (*cont.*)



D. 2. 4. (***Empirical Distribution Function***)
The empirical distribution function is defined as follows:

$$F(x) := \begin{cases} 0 & \text{for} \quad x \le a_1 \\ \sum_{i=1}^{j} f(a_i) & \text{for} \quad a_j < x \le a_{j+1} \\ 1 & \text{for} \quad x > a_k \end{cases}$$

R. 2. 5. (***Some Important Properties of the Distribution Function***)

1.
$$F(x) = P\left(X < x\right) \quad (P: \text{ proportion})$$
2.
$$0 \le F(x) \le 1$$
3.
$$\forall\, x_1, x_2 : x_1 < x_2 \;\Rightarrow\; F\left(x_1\right) \le F\left(x_2\right)$$
4.
$$P(x_1 \le X < x_2) = F\left(x_2\right) - F\left(x_1\right)$$
5.

$F(x)$ is at least left-sided continuous and has at most a finite number of jump discontinuities.

6.

$$x \to -\infty \;\Rightarrow\; F(x) \to 0$$

$$x \to +\infty \;\Rightarrow\; F(x) \to 1$$

R. 2. 6.
A distribution function can be represented

1. in analytic form
2. in tabular form
3. as a graph.

Ex. 2. 1. (*cont.*)

1. *Analytic form*

$$F(x) = \begin{cases} 0 & -\infty < x \le 14.00 \\ 0.033 & 14.0 < x \le 14.70 \\ 0.066 & 14.70 < x \le 15.00 \\ 0.133 & 15.00 < x \le 15.15 \\ 0.200 & 15.15 < x \le 15.60 \\ 0.433 & 15.60 < x \le 16.20 \\ 0.566 & 16.20 < x \le 16.55 \\ 0.666 & 16.55 < x \le 17.05 \\ 0.766 & 17.05 < x \le 17.50 \\ 0.833 & 17.50 < x \le 17.80 \\ 0.900 & 17.80 < x \le 18.00 \\ 0.933 & 18.00 < x \le 18.30 \\ 1.00 & 18.30 < x < +\infty \end{cases}$$

2. Tabular form

Interval	$F(x)$
$]-\infty,\ \ 14.000]$	0.000
$]14.000,\ 14.700]$	0.033
$]14.700,\ 15.000]$	0.066
$]15.000,\ 15.150]$	0.133
$]15.150,\ 15.600]$	0.200
$]15.600,\ 16.200]$	0.433
$]16.200,\ 16.550]$	0.566
$]16.550,\ 17.050]$	0.666
$]17.050,\ 17.500]$	0.766
$]17.500,\ 17.800]$	0.833
$]17.800,\ 18.000]$	0.900
$]18.000,\ 18.300]$	0.933
$]18.300,\ \ +\infty\ [$	1.000

Ex. 2. 1. (*cont.*)
Calculate and interpret

1. $F(15.75)$
2. $F(17.18) - F(15.12)$.

Solution:

1.

$$F(15.75) = 0.433.$$

43.3% of the workers earn less than 15.75 € an hour.

2.

$$F(17.18) - F(15.12) = 0.766 - 0.133 = 0.633.$$

63.3% of the workers earn at least 15.12 € but less than 17.18 € an hour.

R. 2. 6.
The process of dividing the data into different *groups* (viz. *classes*) which are homogeneous within but heterogeneous between themselves is called a *classification*.

D. 2. 5. (***Absolute and Relative Class Frequencies***)
Let the data be divided into the classes C_i, $i = 1, 2, ..., p$.
The *absolute frequency of the class* C_i is defined as follows:

$$F_i := \text{“Number of observations in } C_i\text{”}, \quad i = 1, 2, ..., p$$

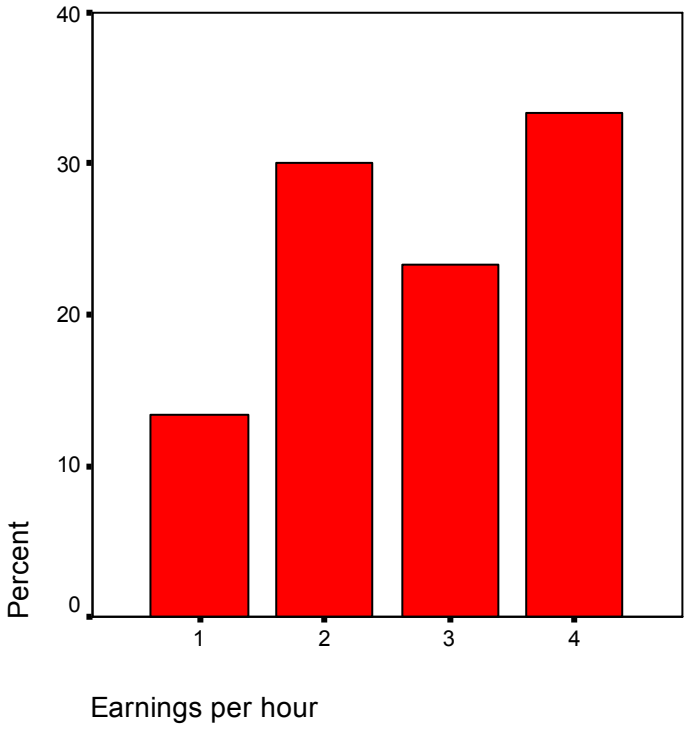
The *relative frequency of the class* C_i is defined as follows:

$$f_i := \frac{1}{n} \cdot H_i, \quad i = 1, 2, ..., p .$$

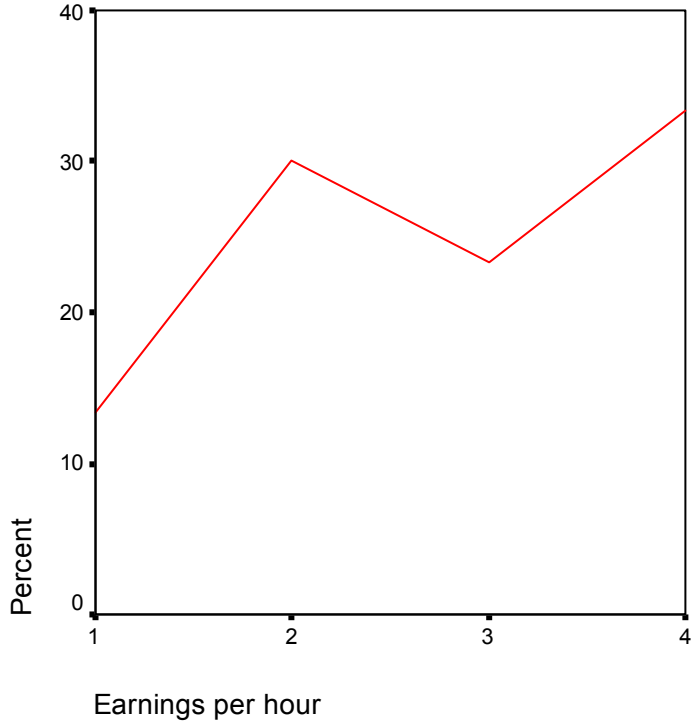
Ex. 2. 1. (*cont.*)

i	C_i	F_i	f_i	$\sum_{j=1}^{i} f_j$	w_i	h_i	m_i
1	[14.00, 15.00[	2	0.067	0.067	1.00	0.067	14.50
2	[15.00, 16.00[	11	0.367	0.434	1.00	0.367	15.50
3	[16.00 17.00[	7	0.233	0.667	1.00	0.233	16.50
4	[17.00 18.40[	10	0.333	1.000	1.40	0.238	17.70
Total		30	1.000				

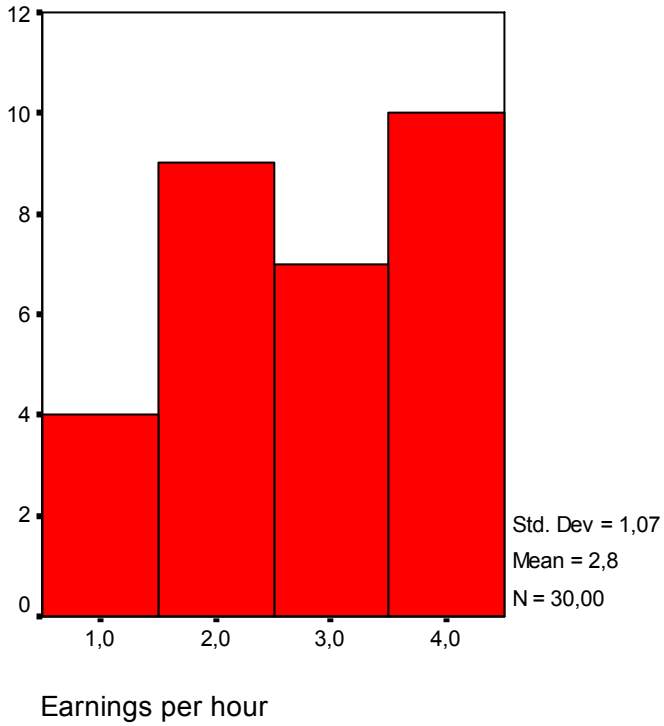
R. 2. 6. (***Graphical Representation of Class Frequencies***)
1. Bar charts



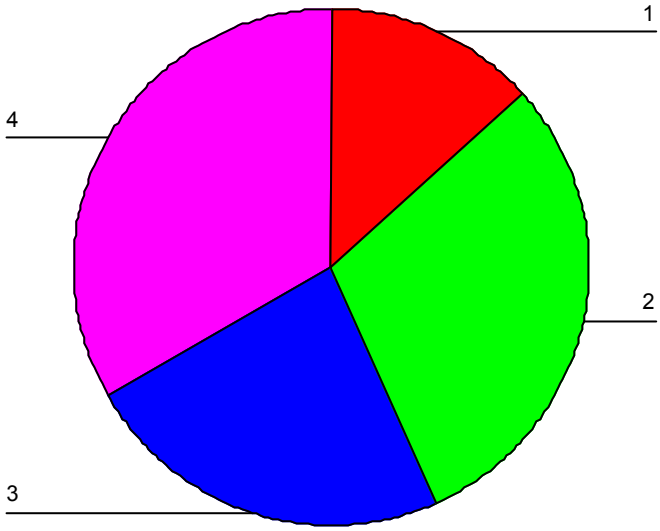
2. Polygon



3. Histogramm (not normed) :



4. Pie chart



R. 2. 7. (***Empirical Distribution of Grouped Data***)

$$F(x) := \begin{cases} 0 & \text{for} \quad x \le m_1 \\ \sum_{i=1}^{j} f(a_i) & \text{for} \quad m_j < x \le m_{j+1} \\ 1 & \text{for} \quad x > m_k \end{cases}$$

Ex. 2. 1. (*cont.*)

$$F(x) = \begin{cases} 0.000 & \text{for} \quad -\infty < x \le 14.50 \\ 0.067 & \text{for} \quad 14.50 < x \le 15.50 \\ 0.434 & \text{for} \quad 15.50 < x \le 16.50 \\ 0.667 & \text{for} \quad 16.50 < x \le 17.70 \\ 1.000 & \text{for} \quad 17.70 < x < +\infty \end{cases}$$

Ex. 2. 1. (*cont.*)
Calculate and interpret

1. $F(16.50)$
2. $F(17.40) - F(15.60)$.

Solution:

1.

$$F(16.50) = 0.434 .$$

About 43.4% of the workers earn less than 16.50 €.

2.

$$F(17.40) - F(15.60) = 0.667 - 0.434 = 0.233 .$$

About 23.3% of the workers earn at least 15.60 € but less than 17.40 € an hour.

(Last revised: 30.03.09)