Kapitel II

Verteilungsanalyse

D. 2. 1. (*Urliste*)
Eine *Urliste* oder *Beobachtungsreihe* ist die Gesamtheit aller Ausprägungen eines Merkmals in der Reihenfolge ihrer Erhebung.

D. 2. 2. (*Absolute und relative Häufigkeit*)
Betrachtet sei ein Merkmal X, das in k Ausprägungen $a_1, a_2, ..., a_k$ vorkommt. Als *absolute Häufigkeit* der Ausprägung a_j, $j = 1, 2, ..., k$, bezeichnet man:

$$H(a_j) := \text{“Anzahl der Fälle, in denen } a_j \text{ auftritt“}, \quad j = 1, 2, ..., k.$$

Als *relative Häufigkeit* der Ausprägung a_j, $j = 1, 2, ..., k$, bezeichnet man:

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

B. 2. 1.
Es gilt:

$$0 \le H(a_j) \le n, \quad j = 1, 2, ..., k,$$

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

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

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

BS. 2. 1.
Für die 30 Akkordarbeiter eines mittleren Betriebes ergab sich folgende Urliste der Stundenlöhne in €

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.

Sei

X : „Stundenlohn in €“.

Hier ist

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

a_j	$H(a_j)$	$h(a_j)$
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
Summe	30	1.000

D. 2. 3. (***Absolute und relative Summenhäufigkeit***)

Unter der *absoluten Summenhäufigkeit* der Ausprägung a_j, $j = 1,2,...,k$, versteht man:

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

Unter der *relativen Summenhäufigkeit* der Ausprägung a_j, $j = 1,2,...,k$, versteht man:

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

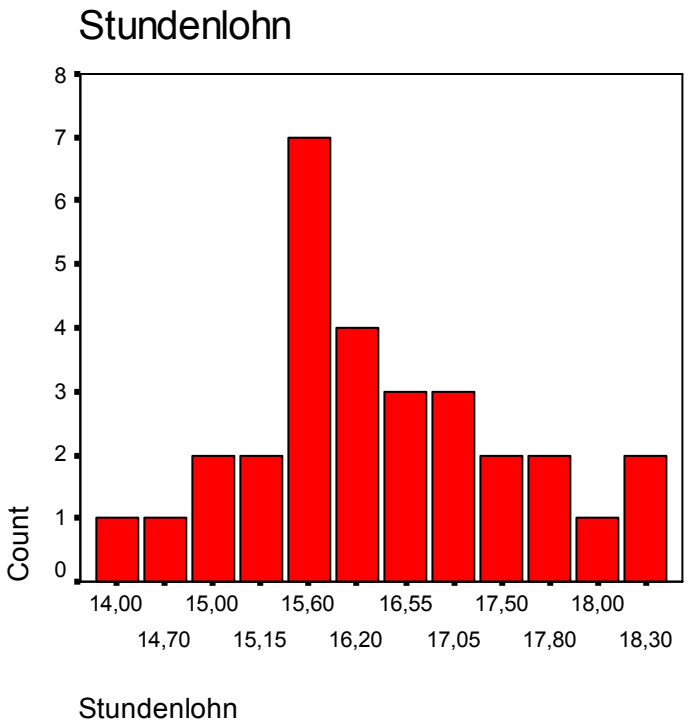
BS. 2. 1. *(Fortsetzung)*

j	a_j	$H(a_j)$	$h(a_j)$	$\sum_{i=1}^{j} H(a_i)$	$\sum_{i=1}^{j} h(a_i)$
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
	Summe	30	1.000		

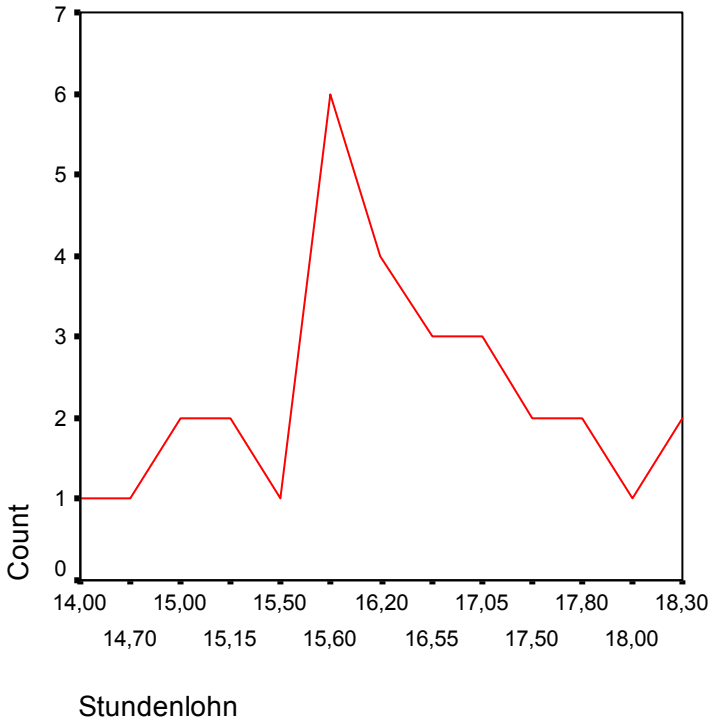
Die Zahl 23 in der Spalte 5 bedeutet: 23 Arbeiter haben einen Stundenlohn von höchstens 17.05 €.
Die Zahl 0.766 in der letzten Spalte bedeutet: Etwa 77 % der Arbeiter haben einen Stundenlohn von höchstens 17.05 €.

B. 2. 2. (***Graphische Darstellung der Häufigkeiten***)
Zur graphischen Darstellung der absoluten und relativen Häufigkeiten gibt es u. a. folgende Möglichkeiten:

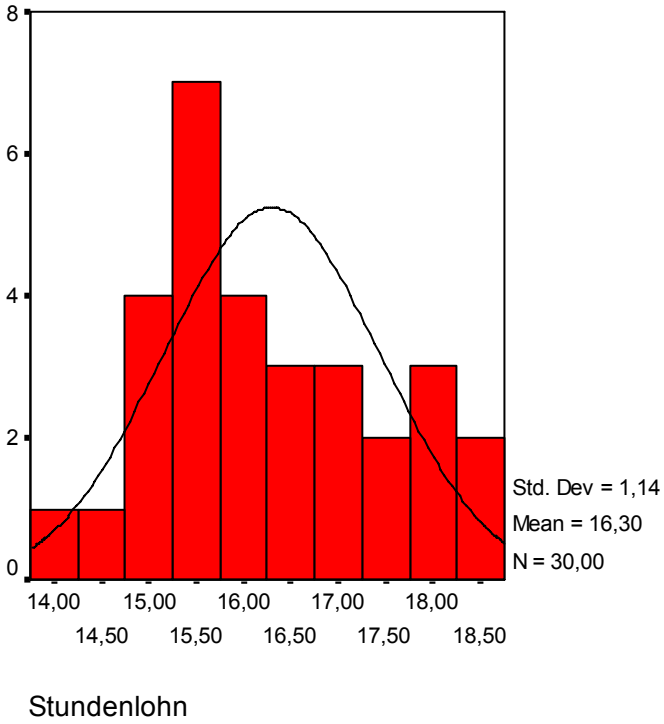
1. Stabdiagramm



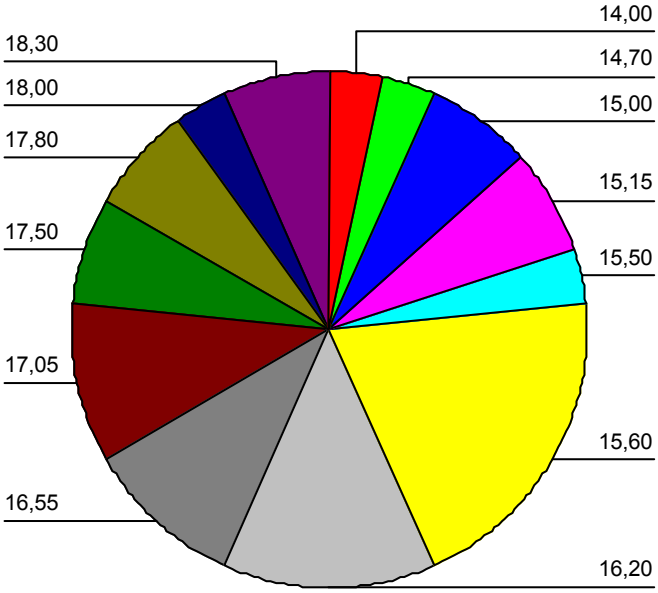
2. Häufigkeitspolygon



3. Histogramm



4. Kreisdiagramm.



<u>D. 2. 4.</u> (*Empirische Verteilungsfunktion*)
Als *empirische Verteilungsfunktion* (bzw. *Summenhäufigkeitsfunktion*) bezeichnet man die Funktion:

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

<u>B. 2. 3</u>. (*Wichtige Eigenschaften der Verteilungsfunktion*)
Es gilt

1.
$$F(x) = A(X < x).$$

2.
$$0 \le F(x) \le 1$$

3.
$$\forall\, x_1, x_2 : x_1 < x_2 \;\Rightarrow\; F(x_1) \le F(x_2)$$

4.
$$A(x_1 \le X < x_2) = F(x_2) - F(x_1)$$

5.

$$x \to -\infty \quad F(x) \to 0$$

$$x \to +\infty \quad F(x) \to 1.$$

6.
Eine empirische Verteilungsfunktion ist mindestens linksseitig stetig und hat höchstens endlich viele Sprungstellen.

BS. 2. 1. *(Fortsetzung)*

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

D. 2. 5. (***Absolute und relative Klassenhäufigkeiten***)
Sei *p* die Anzahl der gebildeten Klassen. Als *absolute Häufigkeit der Klasse* $K_i,\ i = 1,\ 2,...,p$, bezeichnet man:

H_i :="Anzahl der Beobachtungswerte in der Klasse K_i", $\quad i = 1,\ 2,...,p$.

Unter der *relativen Häufigkeit der Klasse* $K_i,\ i = 1,\ 2,...,p$, versteht man:

$$h_i := \frac{H_i}{n},\ i = 1,\ 2,...,p.$$

Dabei ist n die Anzahl der Beobachtungswerte.

B. 2. 4:
Bei der Darstellung der empirischen Verteilungsfunktion im Falle eines gruppierten Datenmaterials werden die Klassenmitten als Repräsentanten der einzelnen Klassen genommen:

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

BS. 2. 1. *(Fortsetzung)*
Wir bilden folgende vier „Lohngruppen“:

$$K_1 = [14.00,\ 15.00[\,,$$
$$K_2 = [15.00,\ 16.00[\,,$$
$$K_3 = [16.00,\ 17.00[\,,$$
$$K_4 = [17.00,\ 18.40[\,.$$

Die Darstellung der Häufigkeiten eines gruppierten Datenmaterials als Histogramm sollte „flächentreu“ sein. Dies bedeutet, dass der Flächeninhalt der einzelnen Säulen gleich den entsprechenden Häufigkeiten sein soll. Damit wird der Gesamtflächeninhalt des Histogramms gleich 1 sein.

Arbeitstabelle

i	K_i	H_i	h_i	$\sum_{j=1}^{i} h_j$	b_i	$l_i := \frac{h_i}{b_i}$	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	1.00				

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

B. 2. 5.
Das Ziel der Klassenbildung besteht darin, die Struktur der untersuchten Gesamtheit deutlich herauszufinden. Hinsichtlich der Zahl der Klassen lässt sich keine generelle Regel formulieren. Werden zu viele Klassen gebildet, bleibt die Struktur unübersichtlich, weil zahlreiche Klassen nur gering oder überhaupt nicht besetzt sind. Bei zu wenigen Klassen kann die charakteristische Form der Verteilung verborgen bleiben.

Die Klasseneinteilung sollte durch das Untersuchungsziel bestimmt werden. Will man beispielsweise zwei Verteilungen miteinander vergleichen, bietet es sich an, für jede Verteilung gleiche Klassen zu bilden.

Als formale Kriterien werden u. a. genannt:

- Die Anzahl der Klassen sollte zwischen 10 und 20 liegen. Für kleinere Datensätze wird als Faustregel verwendet: $p \leq \sqrt{n}$.
- Möglist gleiche Klassenbreiten wählen.
- Der in der Gesamtheit am häufigsten vorkommende Merkmalswert sollte die Klassenmitte der Klasse mit der größten Besetzung bilden
- Die Klassenmüssen eindeutig voneinander abgegrenzt werden und lückenlos aufeinander folgen.
- Interessiert nur ein bestimmter Bereich der erhobenen Daten, kann es sinnvoll sein, offene Randklassen zu bilden.

(Letzte Aktualisierung: 23.09.12)