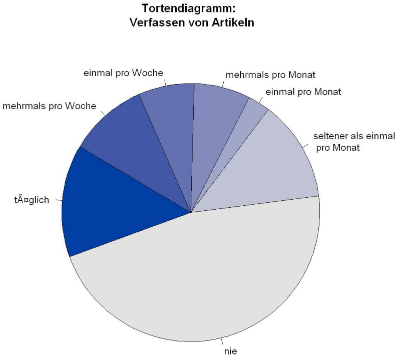
Kreisdiagramm, Tortendiagramm

Darstellung der relativen (absoluten) Häufigkeiten als Fläche eines Kreises

Anwendung:

- Nominale Merkmale
- Ordinale Merkmale (Problem: Ordnung nicht korrekt wiedergegeben)
- Gruppierte Daten

Beispiel Redakteure: Kreisdiagramm



Stabdiagramm, Säulen- und Balkendiagramm

- *Stabdiagramm*:
  Trage über $a_1, \ldots, a_k$ jeweils einen zur x-Achse senkrechten Strich (Stab) mit Höhe $h_1, \ldots, h_k$ (oder $f_1, \ldots, f_k$) ab.
- *Säulendiagramm*:
  wie Stabdiagramm, aber mit Rechtecken statt Strichen.
- *Balkendiagramm*:
  wie Säulendiagramm, aber mit vertikal statt horizontal gelegter x-Achse.

Säulendiagramm

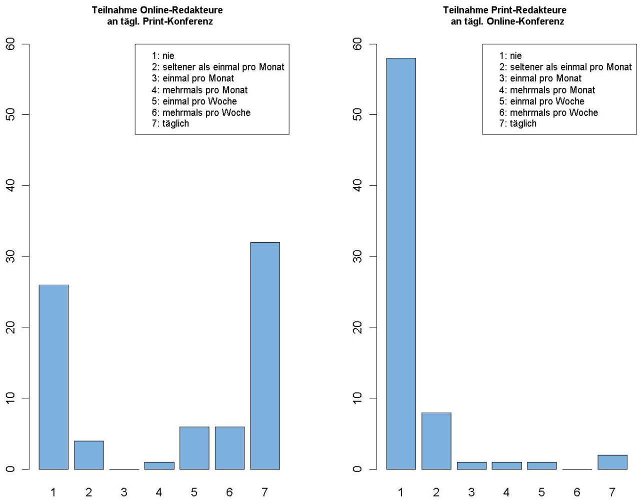
Darstellung der absoluten oder relativen Häufigkeiten als Höhen (Längen)

x-Achse: Ausprägungen des Merkmals
y-Achse: absolute/ relative Häufigkeiten

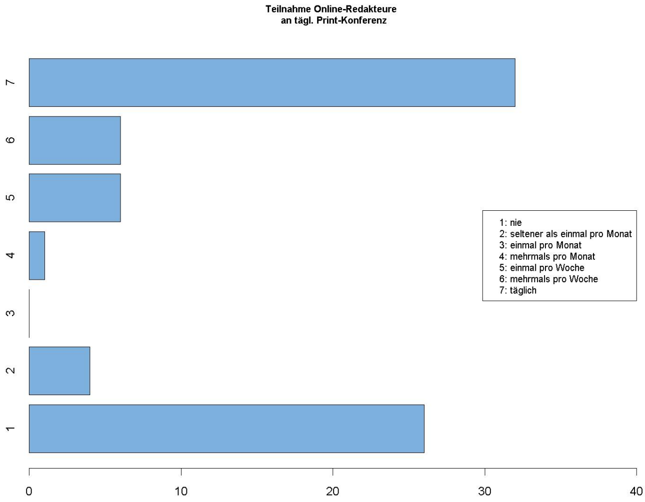
Anwendungen:

- Ordinale Merkmale
- Metrische Merkmale mit wenigen Ausprägungen
- Nominale Merkmale (Problem: Ordnung nicht vorhanden)

Beispiel Redakteure: Säulendiagramm vertikal



Beispiel Redakteure: Säulendiagramm horizontal



Stapeldiagramm

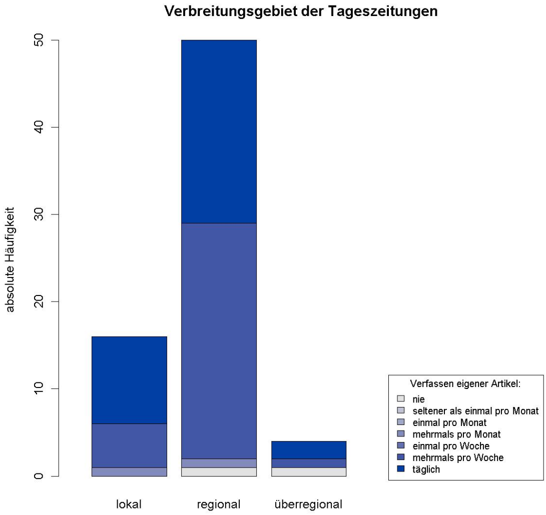
Darstellen der absoluten oder relativen Häufigkeiten als Länge. Die Abschnitte werden übereinander in verschiedenen Farben gestapelt.

Anwendungen:

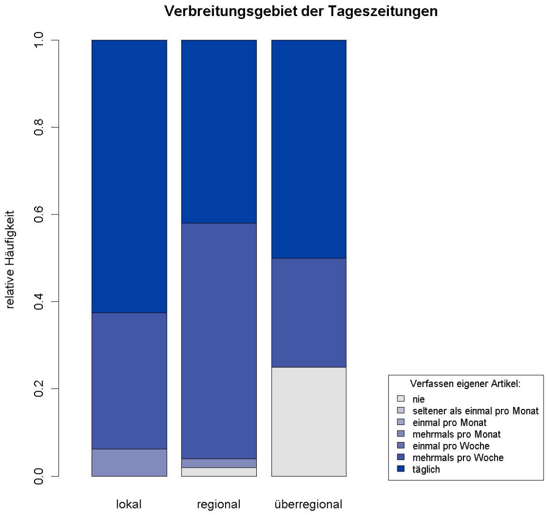
- Ordinale Daten
- Gruppierte Daten
- Metrische Daten mit wenigen Ausprägungen

Besonders geeignet für den Vergleich verschiedener Gruppen durch nebeneinander liegende Stapel. Zu beachten ist dann die Unterscheidung: relative Häufigkeit $\leftrightarrow$ absolute Häufigkeit

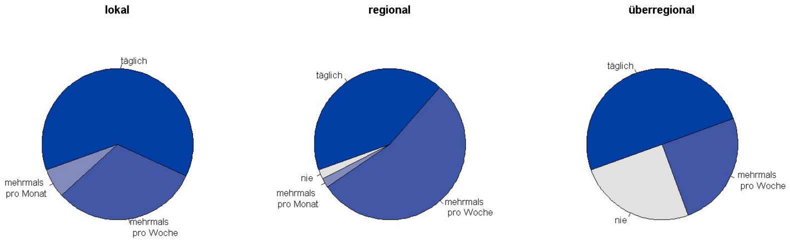
Beispiel Redakteure: Stapeldiagramm I



Beispiel Redakteure: Stapeldiagramm II



Beispiel Redakteure: Vergleich mit Kreisdiagramm



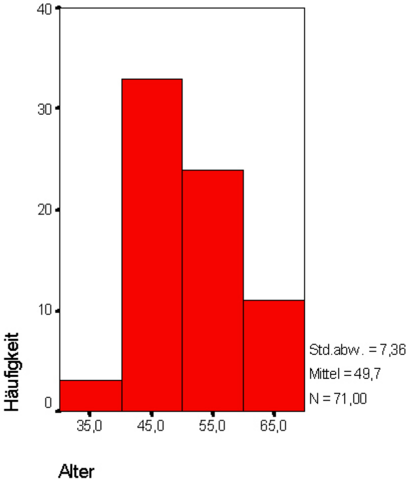
Das Histogramm

Darstellung der relativen Häufigkeiten durch Flächen
(Prinzip der Flächentreue)

Vorgehen:

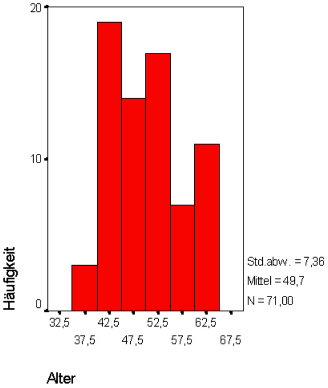
1. Aufteilung in Klassen (falls die Daten noch nicht gruppiert sind)
2. Bestimmung der relativen Häufigkeiten $f_j = \frac{n_j}{n}$
3. Bestimmung der Höhen h_j, so dass gilt $b_j \cdot h_j = f_j$
   wobei b_j: Breite der Klasse j.

Beispiel: Alter der Redakteure

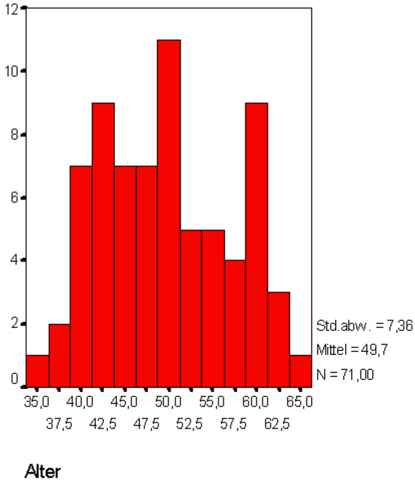


Beispiel: Alter der Redakteure

Altersklassen in Abständen von 5 Jahren



Histogramm mit Standardeinstellung aus SPSS



Histogramm

- Anwendung bei metrischen Daten
- Beachte: Abhängigkeit von der Breite
- Klasse inhaltlich vorgeben, verschiedene Varianten ansehen.
- Vorsicht bei Rändern

Stamm-Blätter-Diagramm

(Stem and leaf plot)

Spezielles Histogramm mit

- Klassen nach Dezimalsystem
- Einzeldaten reproduzierbar

Beispiel: Alter der Redakteure

Alter Stem-and-Leaf Plot

Frequency	Stem & Leaf
3,00	3 . 678
19,00	4 . 0011111222222233444
14,00	4 . 56667788888999
17,00	5 . 00000011222224444
7,00	5 . 6788899
11,00	6 . 00000112234

Stem width: 10

Each leaf: 1 case(s)

Empirische Verteilungsfunktion

$$H(x) := \text{Anzahl der Werte} <= x$$

$$F(x) = H(x)/n = \text{Anteil der Werte } x_i \text{ mit } x_i \leq x$$

bzw.

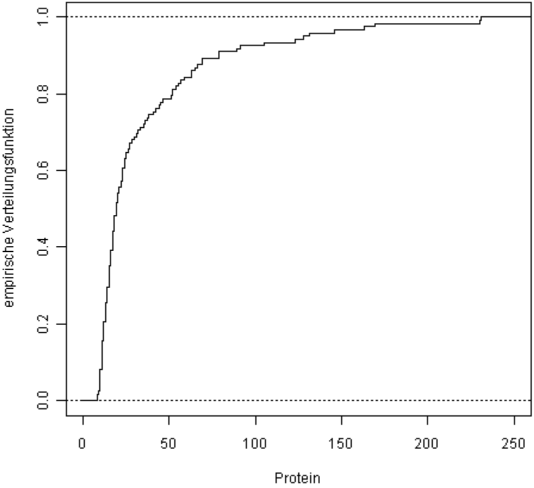
$$F(x) = f(a_1) + \ldots + f(a_j) = \sum_{i:a_i \leq x} f_i,$$

wobei $a_j \leq x$ und $a_{j+1} > x$ ist.

Eigenschaften von $F(x)$

- monoton wachsende Treppenfunktionen mit Sprüngen an den Ausprägungen $a_1, \ldots, a_k$
- Sprunghöhen: $h_1, \ldots, h_k$ bzw. $f_1, \ldots, f_k$
- rechtsseitig stetig
- $H(x) = 0$ für $x < a_1$, $H(x) = n$ für $x \geq a_k$
  $F(x) = 0$ für $x < a_1$, $F(x) = 1$ für $x \geq a_k$

Beispiel für eine Empirische Verteilungsfunktion



Eindimensionale statistische Kennwerte

Lagemaßzahlen

- Wo liegt die Masse der Daten?
- Wo liegt die Mehrzahl der Daten?
- Wo liegt die Mitte der Daten?
- Welche Merkmalsausprägung ist typisch für die Häufigkeitsverteilung?

Statistische Kennwerte

Streumaßzahlen

- Über welchen Bereich erstrecken sich die Daten?
- Wie groß ist die Schwankung der Ausprägungen?

Der Modus

Definition: Häufigster Wert

Eigenschaften:

- oft nicht eindeutig
- nur bei gruppierten Daten oder bei Merkmalen mit wenigen Ausprägungen sinnvoll
- stabil bei allen eindeutigen Transformationen
- geeignet für alle Skalenniveaus

Der Median

Definition: Wert für den gilt

Mindestens 50% der Daten sind kleiner oder gleich *med*
Mindestens 50% der Daten sind größer oder gleich *med*

$$med = \begin{cases} x_{(k)} & \text{falls } k = \frac{n+1}{2} \text{ ganze Zahl} \\ \frac{1}{2}(x_{(k)} + x_{(k+1)}) & \text{falls } k = \frac{n}{2} \text{ ganze Zahl} \end{cases}$$

$x_{(1)}, \ldots, x_{(n)}$ sind geordnete Werte

Eigenschaften des Medians

- anschaulich
- stabil gegenüber monotonen Transformationen
- geeignet für ordinale Daten
- stabil gegenüber Ausreißern

Das Quantil

Definition: Wert für den gilt:

Mindestens Anteil p der Daten sind kleiner oder gleich x_p
Mindestens Anteil $1-p$ der Daten sind größer oder gleich x_p

$$\begin{cases} x_{(k)} & \text{falls } np \text{ keine ganze Zahl und } k \text{ kleinste Zahl} > np \\ \in \left[x_{(k)}; x_{(k+1)}\right] & \text{falls } k = np \text{ ganze Zahl} \end{cases}$$

Es gibt weitere Definitionen von Quantilen (in R 9 Typen), die sich aber in der Praxis kaum unterschieden.

Boxplot

Einfacher Boxplot

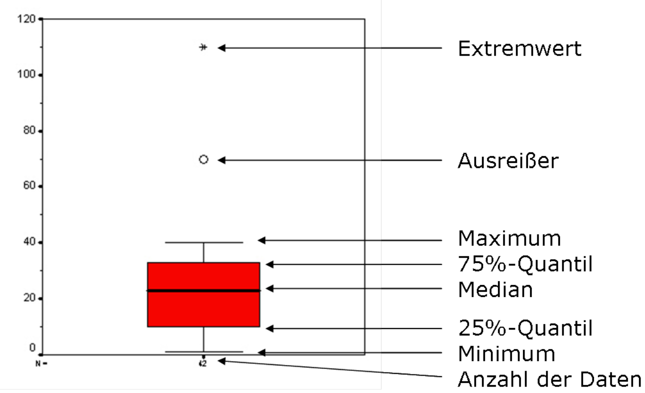
- $\tilde{x}_{0.25}$ = Anfang der Schachtel (Box)
  $\tilde{x}_{0.75}$ = Ende der Schachtel
  d_Q = Länge der Schachtel
- Der Median wird durch den Strich in der Box markiert
- Zwei Linien („whiskers“) außerhalb der Box gehen bis zu x_{min} und x_{max}.

Modifizierter Boxplot

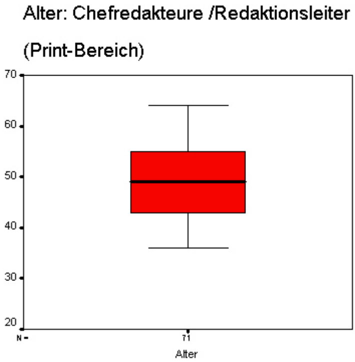
Die Linien außerhalb der Schachtel werden nur bis zu x_{min} bzw. x_{max} gezogen, falls x_{min} und x_{max} innerhalb des Bereichs $[z_u, z_o]$ der Zäune liegen.
$z_u = \tilde{x}_{0.25} - 1,5 d_Q$, $z_o = \tilde{x}_{0.75} + 1,5 d_Q$
Ansonsten gehen die Linien nur bis zum kleinsten bzw. größten Wert innerhalb der Zäune, die außerhalb liegenden Werte werden individuell eingezeichnet.

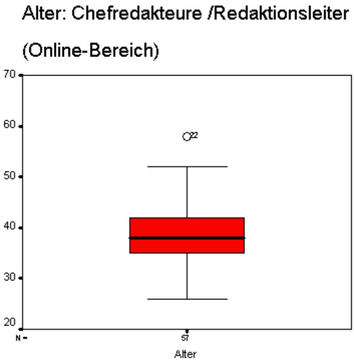
Boxplot

- Eindimensionale Darstellung auf der zugehörigen Skala
- Visualisieren der 5-Punkte-Zusammenfassung

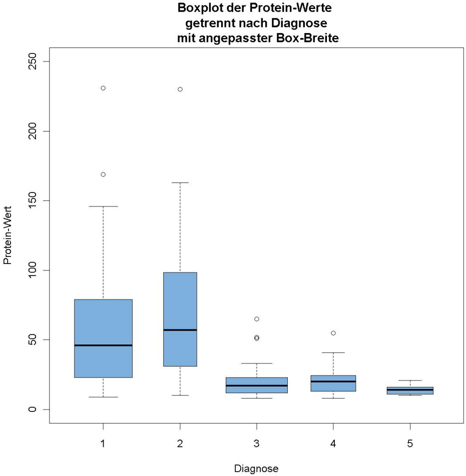


Beispiel Redakteure: Boxplot





Beispiel: Boxplot für Gruppen



Boxplot: Vor- und Nachteile

pro:

- kompakt
- geeignet für Vergleiche
- Ausreißer sichtbar
- Schiefe sichtbar

contra

- gegen Intuition (Viel Farbe – wenig Daten)
- Bimodale Verteilungen nicht sichtbar
- Ausreißer sichtbar
- Breite redundant