

2. Univariate deskriptive Statistik

- ▶ Die univariate deskriptive Statistik stellt Instrumente zur Beschreibung von eindimensionalen Daten bereit.
- ▶ Insbesondere sind dies Methoden/Maßzahlen, die
 - ▶ das Datenmaterial strukturieren.
 - ▶ Beobachtungen grafisch aufarbeiten.
 - ▶ eine zentrale Lage von Beobachtungen ausdrücken.
 - ▶ eine Streuung in den Beobachtungen erfassen.
 - ▶ die Form bzw. das Profil der (Häufigkeits-) Verteilung beschreiben.

- ▶ Ausgangspunkt: Beobachtung oder Messung eines Merkmals X an n statistischen Einheiten
- ▶ Menge der Ausprägungen, die das Merkmal X annimmt oder annehmen kann: $\{a_1, \dots, a_k\}$
- ▶ Ergebnis der i -ten Beobachtung/Messung:
 $x_i \in \{a_1, \dots, a_k\}$
- ▶ nach der Beobachtung/Messung an n statistischen Einheiten: Daten, Urliste, Beobachtungsreihe, Stichprobe x_1, \dots, x_n

i	Maßkrug Bier	Geschlecht	Einstiegs- gehalt	Noten im Studium
1	1	männlich	45 000	gut
2	0	weiblich	46 000	gut
3	3	männlich	38 000	schlecht
4	4	männlich	42 000	mittel
5	4	weiblich	47 000	mittel
6	2	weiblich	42 000	gut
7	0	weiblich	41 000	gut
8	3	männlich	45 000	schlecht
9	0	männlich	40 000	mittel
10	5	männlich	42 142	mittel



- ▶ **absolute Häufigkeit** der Ausprägung a_j , d. h. Anzahl der x_i aus x_1, \dots, x_n mit $x_i = a_j$:

$$h_j = h(a_j)$$

- ▶ Die Summe der absoluten Häufigkeiten ergibt die Gesamtanzahl n der Beobachtungen.
- ▶ **relative Häufigkeit** der Ausprägung a_j :

$$f_j = f(a_j) = h(a_j) / n$$

- ▶ Die relativen Häufigkeiten summieren sich zu 1 auf.

▶ **absolute Häufigkeitsverteilung:**

Merkmalsausprägungen a_1, \dots, a_k zusammen mit den zugehörigen absoluten Häufigkeiten h_1, \dots, h_k

▶ **relative Häufigkeitsverteilung:**

Merkmalsausprägungen a_1, \dots, a_k zusammen mit den zugehörigen relativen Häufigkeiten f_1, \dots, f_k

▶ klassische Form der Darstellung ist die Tabelle



- ▶ Visualisierung der Verteilung
 - ▶ diskreter Merkmale, wenn Anzahl k der Merkmalsausprägungen klein:
Stab- und Kreisdiagramme
 - ▶ zumindest ordinal skaliertes Merkmale, wenn Anzahl k der Merkmalsausprägungen groß:
Histogramme

- ▶ **Stabdiagramm:** höhenproportionale Darstellungsform einer Häufigkeitsverteilung
 - ▶ Trage über a_1, \dots, a_k jeweils einen zur x -Achse senkrechten Strich (Stab) mit Höhe h_1, \dots, h_k (oder f_1, \dots, f_k) ab.
- ▶ **Säulendiagramm:** wie Stabdiagramm, aber mit Rechtecken statt Strichen (Breite der Rechtecke bedeutungslos)



Kandidat	Stimmenanteil
Bengler	10.2 %
Hauer	40.5 %
Mooser-Niefanger	23.2 %
Schneider	26.1 %

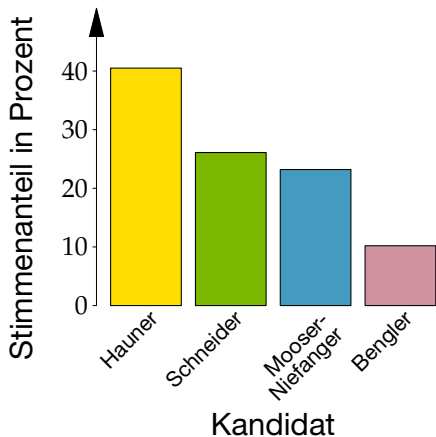
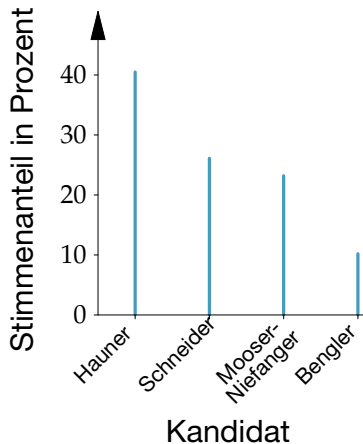


Abb.: Stimmenverteilung bei der Landratswahl Freising, 2014.
Stab- (links) und Säulendiagramm (rechts)

- ▶ **Kreisdiagramm:** flächenproportionale Darstellungsform einer Häufigkeitsverteilung
 - ▶ Zu jeder Merkmalsausprägung a_j wird ein Kreissektor gebildet.
 - ▶ Die Flächen der Sektoren werden proportional zu den Häufigkeiten gewählt.
 - ▶ Der Winkel α des Kreissektors j ist $\alpha_j = f_j \cdot 360^\circ$.



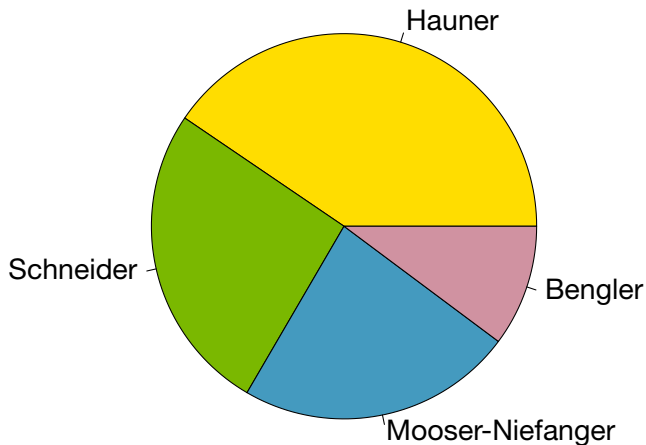
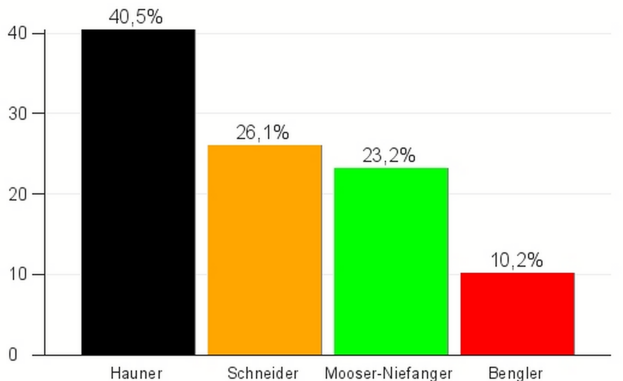


Abb.: Stimmenverteilung bei der Landratswahl Freising, 2014

- ▶ Flächen können in ihren Größenverhältnissen leicht falsch beurteilt werden.
- ▶ Kreisdiagramme deshalb nur bei nominalen Merkmalen einsetzen, sonst Stab-/Säulendiagramme vorziehen.
- ▶ Auch bei nominalen Merkmalen sind Stab-/Säulendiagramme aber möglich.

Landratswahl Freising

Kommunalwahl 16. März 2014



Stand: 16.03.2014 21:08

www.wochenanzeiger.de

Abb.: Stimmenverteilung bei der Landratswahl

- ▶ **Body-Mass-Index (BMI):** Verhältnis von Körpergewicht (kg) und Körpergröße² (m²)

$$\text{BMI} = \frac{\text{Körpergewicht}}{\text{Körpergröße}^2}$$

20.8,	27.1,	35.0,	29.0,	26.7,	30.0,
29.7,	23.3,	17.0,	23.2,	34.9,	24.9,
27.6,	19.5,	22.6,	27.9,	24.6,	28.0,
28.6,	25.0,	25.9,	18.8,	25.6,	21.1,
20.0,	25.8,	30.0,	27.1,	26.7,	22.0,
21.0,	21.6,	23.7,	21.5,	22.1,	25.5,
23.1,	28.7,	21.7,	26.5,	28.8,	24.0,
21.9,	30.0,	26.5,	20.3,	28.1,	26.6,
24.8,	23.3,	18.5,	25.5,	28.8,	24.7,
25.0,	26.6,	24.5,	32.0,	32.2,	28.8

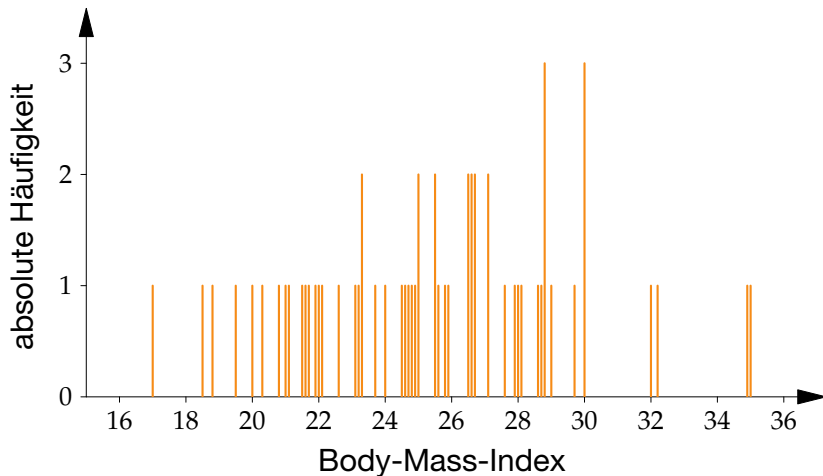


Abb.: Absolute Häufigkeitsverteilung der BMI-Daten ($n = 60$)

- ▶ **Histogramm:** flächenproportionale Darstellungsform einer Häufigkeitsverteilung
 - ▶ Bilde Klassen durch Gruppierung von Merkmalsausprägungen.
 - ▶ Zeichne über den Klassen $[c_0, c_1], (c_1, c_2], (c_2, c_3], \dots, (c_{k-1}, c_k]$ Rechtecke mit
 - ▶ Breite: $d_j = c_j - c_{j-1}, j = 1, \dots, k$
 - ▶ Höhe: proportional zu h_j/d_j bzw. f_j/d_j , sowie
 - ▶ Fläche: proportional zu h_j bzw. f_j .



17.0, 18.5, 18.8, 19.5, 20.0, 20.3,
20.8, 21.0, 21.1, 21.5, 21.6, 21.7,
21.9, 22.0, 22.1, 22.6, 23.1, 23.2,
23.3, 23.3, 23.7, 24.0, 24.5, 24.6,
24.7, 24.8, 24.9, 25.0, 25.0, 25.5,
25.5, 25.6, 25.8, 25.9, 26.5, 26.5,
26.6, 26.6, 26.7, 26.7, 27.1, 27.1,
27.6, 27.9, 28.0, 28.1, 28.6, 28.7,
28.8, 28.8, 28.8, 29.0, 29.7, 30.0,
30.0, 30.0, 32.0, 32.2, 34.9, 35.0



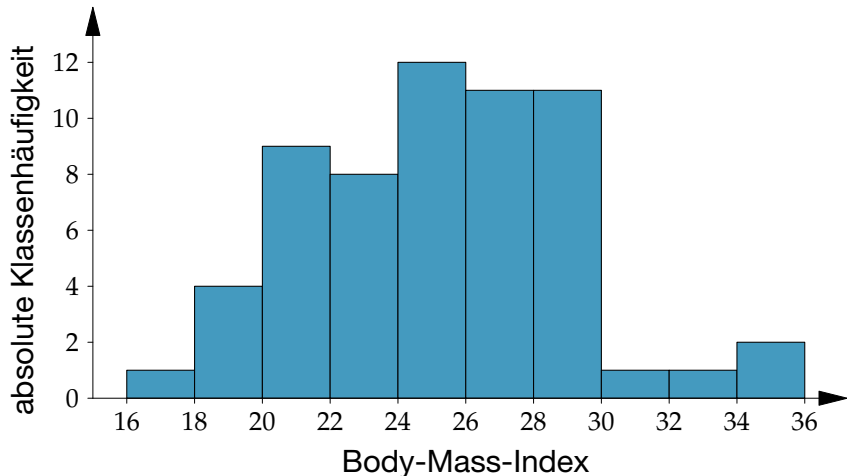


Abb.: Klassierte Häufigkeitsverteilung (gleiche Klassenbreiten) der BMI-Daten ($n = 60$)

- ▶ Gewichtsklassifikation bei Erwachsenen anhand des BMI (nach WHO, 2000):
 - ▶ Untergewicht: $\text{BMI} < 18.5$
 - ▶ Normalgewicht: $18.5 \leq \text{BMI} < 25$
 - ▶ Übergewicht: $\text{BMI} \geq 25$

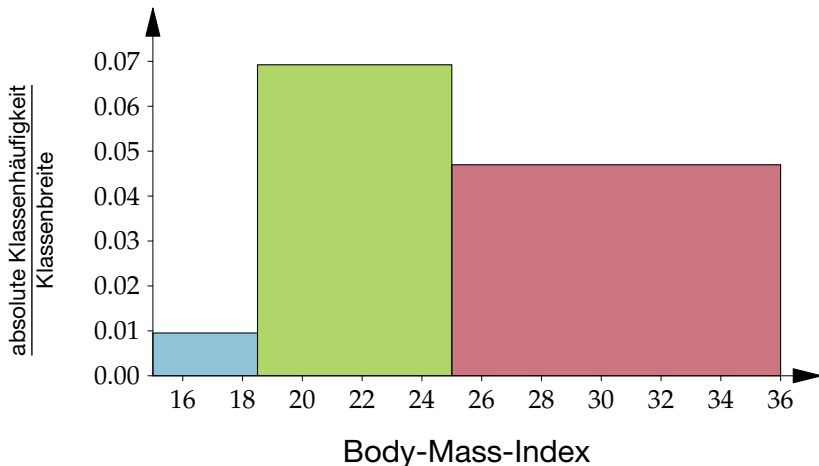


Abb.: Klassierte Häufigkeitsverteilung (unterschiedliche Klassenbreiten) des BMI, $n = 60$

- ▶ Klassengrenzen sollten möglichst einfache Werte sein.
- ▶ Falls möglich und sinnvoll, sollten die Klassenbreiten d_j gleich groß sein.
- ▶ Bei konstanter Klassenbreite ist das Histogramm höhenproportional.
- ▶ Faustregeln für die Anzahl von Klassen sind:
 1. $k \approx \sqrt{n}$
 2. $k \approx 2\sqrt{n}$
 - ⋮

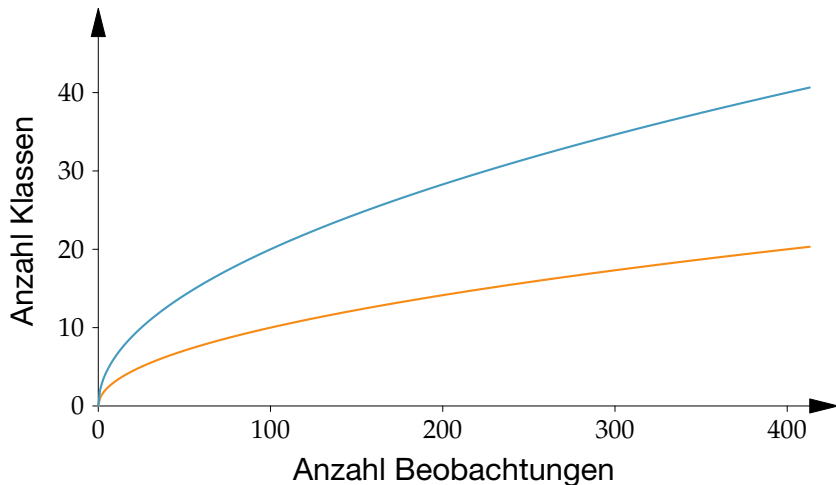


Abb.: Funktionsgraf von \sqrt{n} (orange) und $2\sqrt{n}$ (blau)

- ▶ Faustregel 1: $k \approx \sqrt{n}$, also $\sqrt{60} = 7.7 \approx 10$
- ▶ Faustregel 2: $k \approx 2\sqrt{n}$, also $2\sqrt{60} = 15.5 \approx 20$

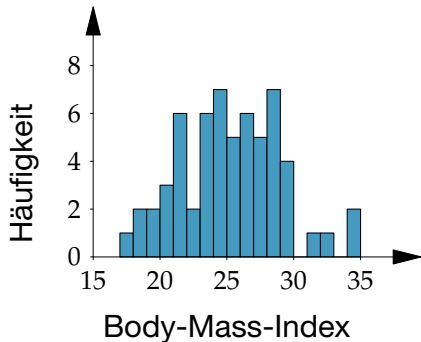
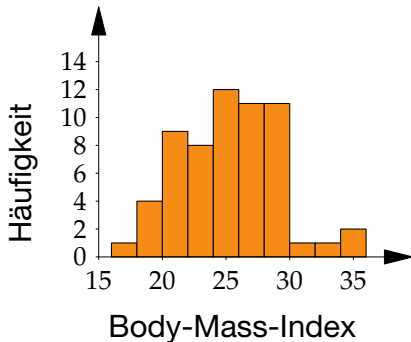


Abb.: Histogramme der BMI-Daten mit 10 Klassen (links) und 20 Klassen (rechts)



- ▶ **unimodale Verteilung:** ein ausgeprägter Gipfel
- ▶ **bimodale Verteilung:** zwei ausgeprägte Gipfel
- ▶ **multimodale Verteilung:** mehrere Gipfel

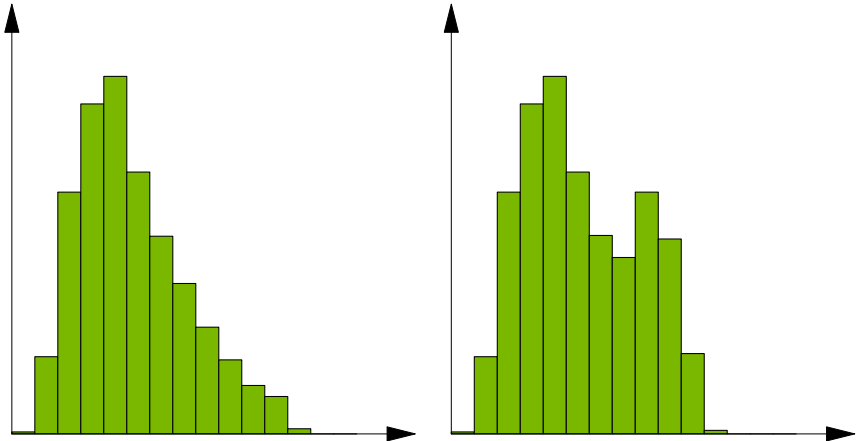


Abb.: Uni- (links) und bimodale (rechts) Verteilung

- ▶ **symmetrische Verteilung:** Es gibt eine Symmetrieachse, so dass die rechte und die linke Hälfte der Verteilung annähernd zueinander spiegelbildlich sind.
- ▶ **unsymmetrische Verteilung:**
 - ▶ **linkssteil** (oder **rechtsschief**): Überwiegender Anteil von Daten ist linksseitig konzentriert.
 - ▶ **rechtssteil** (oder **linksschief**): Überwiegender Anteil von Daten ist rechtsseitig konzentriert.

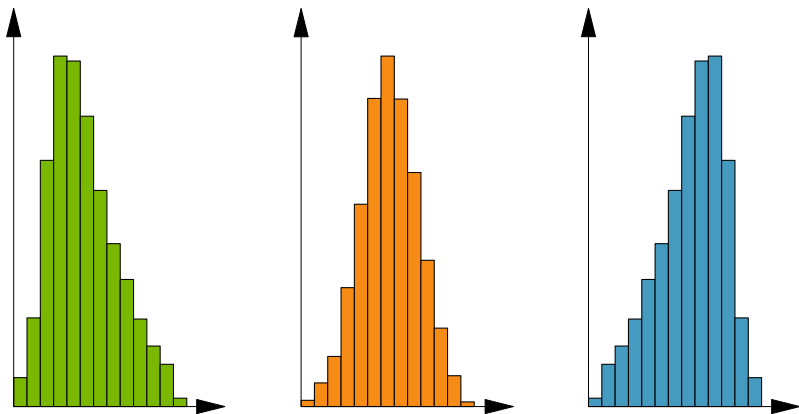


Abb.: Linkssteile, symmetrische und rechtssteile Verteilung



- ▶ Die Beschriftung einer Tabelle steht immer über der Tabelle.
- ▶ Die Abbildungsbeschriftung steht immer unterhalb der Abbildung.