

4. Analyse univariater Daten: Übersicht

Mathematik ist die Wissenschaft der reinen Zahl, Statistik die der empirischen Zahl

Von **univariaten Daten** spricht man, wenn bei der Datenerhebung nur ein Merkmal erfasst wurde bzw. man sich nur für ein Merkmal interessiert. (\Rightarrow die nächsten Wochen).

Interessiert man sich für die Zusammenhänge zwischen mehreren Merkmalen (z.B. Verkehrsdichte *und* Geschwindigkeit), benötigt man die Mittel der **multivariaten** Datenanalyse (\Rightarrow gegen Ende dieser Vorlesung).

Es gibt zwei grundsätzliche Analysemethoden:

- Analyse **nichtklassierter Daten**: Man nimmt die einzelne Elemente der statistischen Reihe direkt als Ausgangspunkt der Untersuchungen.
 - Günstig, falls der Umfang der Datenreihe gering ist oder es nur wenige unterschiedliche Merkmalsausprägungen gibt.
 - Die einzige Analysemethode, falls das zu untersuchende Merkmal nicht kardinalskaliert ist.
- Analyse **klassierter Daten**: Vor der Analyse fasst man jeweils mehrere Merkmalsausprägungen in (Merkmalswerte-)Klassen zusammen.
 - Geeignet, falls es (bei entsprechendem Umfang der Datenreihe) sehr viele verschiedene Merkmalsausprägungen gibt.
 - Vorteil: Übersichtliche Darstellung und neue Analysemethoden, z.B. Darstellung der Häufigkeitsdichte
 - Nachteil: Informationsverlust gegenüber unklassierten Daten.

5: Häufigkeitsverteilung und Summenverteilung

Zehn von hundert Menschen haben Ahnung vom Prozentrechnen. Das sind über 17%

aus einem Zeitungskommentar

5.1. Häufigkeitsverteilung nichtklassierter Daten

Ist das interessierende Merkmal nur nominalskaliert, lassen sich die Daten nur durch eine "Strichliste" aggregieren. Eine Strichliste ist darüber hinaus prinzipiell bei allen diskreten Daten (sowie bei klassierten Daten \Rightarrow Kap. 5.3) erstellbar. Man erhält daraus absolute und relative Häufigkeiten:

Für eine statistische Masse vom Umfang n und ein beliebig skaliertes Merkmal X mit $m \leq n$ verschiedenen Ausprägungen $x_j, j = 1, \dots, m$ heißt die Anzahl

$$h_j = h(X = x_j)$$

der Merkmalsausprägungen x_j die **absolute Häufigkeit**.

Der Anteilswert

$$f(X = x_j) = f_j = \frac{h_j}{n}$$

heißt **relative Häufigkeit**, und die Zusammenstellung der Paare (x_j, h_j) bzw. (x_j, f_j) heißt **absolute und relative Häufigkeitsverteilung**.

Frage: Warum ist diese Definition für stetige Merkmale zwar anwendbar aber sinnlos?

5.2(a): Summenhäufigkeiten nichtklassierter Daten

Ist das Merkmal X mindestens ordinalskaliert, kann man auch Summenhäufigkeiten bilden:

Absolute Summenhäufigkeit:

$$H_j = H(X \leq x_j) = \sum_{j'=1}^j h(X = x_{j'}) = \sum_{j'=1}^j h_{j'}$$

Relative Summenhäufigkeit:

$$F_j = H_j/n.$$

Hierbei sind die Ausprägungen x_j , $j = 1, \dots, m$ des Merkmals X aufsteigend geordnet.

Fragen:

1. Woran sieht man direkt (an einem einzigen Zeichen!), dass die Daten bei den Summenhäufigkeiten mindestens ordinalskaliert sein müssen? am \leq -Zeichen. Die Relation "größer oder gleich" legt ja eine Rangfolge fest. Diese Eigenschaft haben nur mindestens ordinalskalierte Daten.
2. Ermitteln Sie die diversen aggregierten Größen für das Klausurpunkte-Beispiel von Kapitel 3.

5.2(b): Verteilungsfunktion nichtklassierter Daten

Sie wird aus den Summenhäufigkeiten hergeleitet:

Die Funktion

$$F(x) = \begin{cases} 0 & \text{falls } x < x_1, \\ F_j & \text{falls } x_j \leq x < x_{j+1} \\ & \text{mit } j = 1, 2, \dots, m-1, \\ 1 & \text{falls } x > x_m \end{cases}$$

heißt **Empirische Verteilungsfunktion**.

Hierbei sind die m verschiedenen Ausprägungen x_j des Merkmals X aufsteigend geordnet.

Die empirische Verteilungsfunktion ist für beliebige reellwertige Argumente x definiert. An den Punkten x_j springt sie jeweils um den Betrag f_j .

Aufgabe: Ermitteln Sie die absolute und relative Summenfunktion sowie die empirische Verteilungsfunktion für das Klausurbeispiel von Kap. 3.

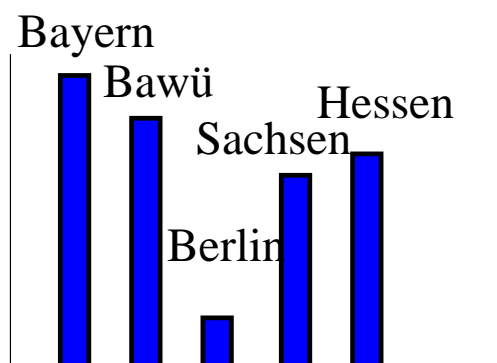
5.3. Darstellung nichtklassierter Daten

Statistik ist wie ein Bikini: Er zeigt das meiste, doch verhüllt das Wesentliche

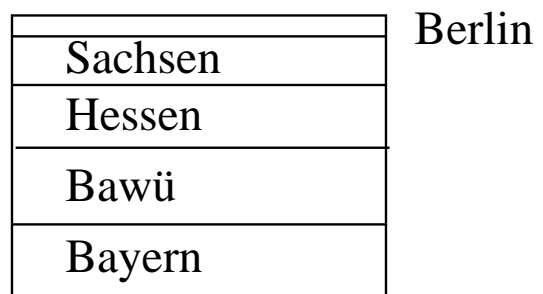
Anonymus

Darstellung der absoluten und relativen Häufigkeiten diskreter, beliebig skaliertes Merkmale mit relativ wenig Merkmalsausprägungen:

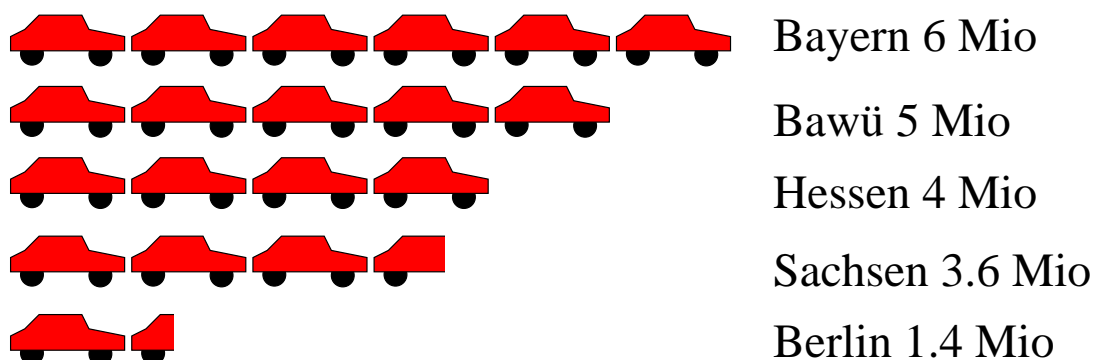
(a) Stabdiagramm



(b) Rechteckdiagramm

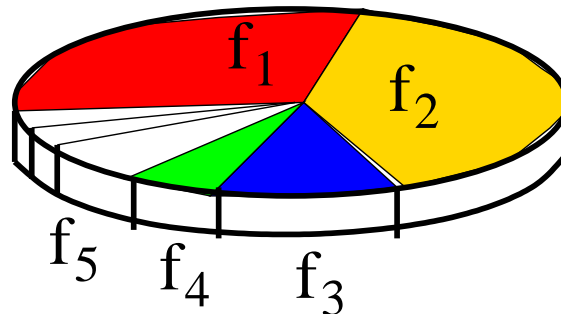


(c) Piktogramm



5.3(b) Darstellung nichtklassierter Daten

(d) Kreis– oder Tortendiagramm

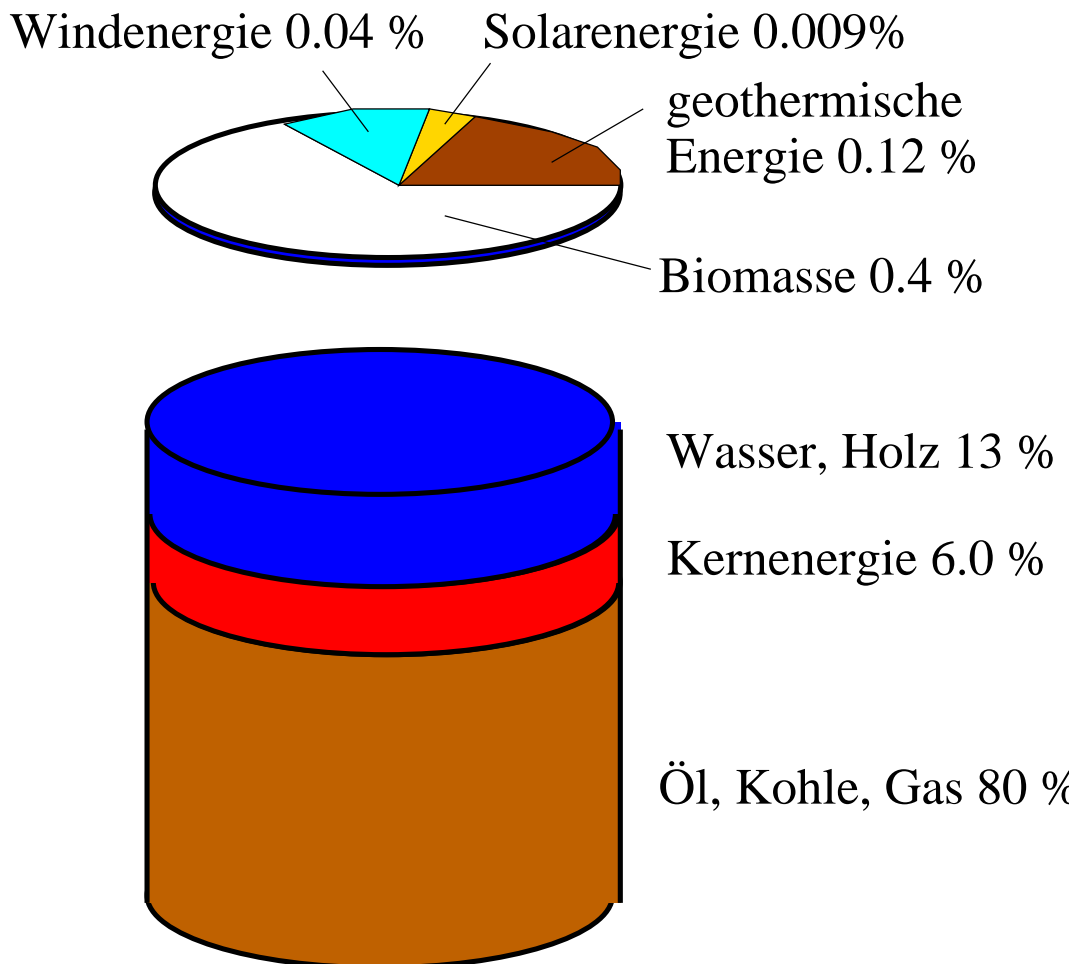


Will man diskreter Merkmale mit *stark unterschiedlichen (relativen) Häufigkeiten* der einzelnen Merkmalsausprägungen anschaulich darstellen, bedient man sich perspektivisch-dreidimensionaler Darstellungen diverser Kombinationen der bisherigen Diagramme:

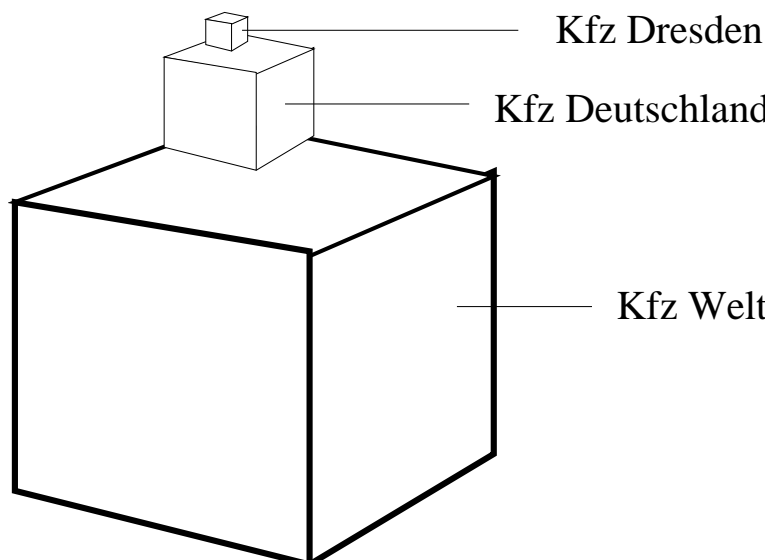
- (e) **3D-Säulendiagramm** als Kombination von Rechteckdiagramm und Kreis- bzw. Tortendiagramm, oder
- (f) **Würfeldiagramm** als dreidimensionale Verallgemeinerung des Rechteckdiagramms.
- (g) Zwei- oder dreidimensional gezeichnete, unterschiedlich große Piktogramme sind ebenfalls möglich.

5.3(c) Dreidimensionale Darstellungen

3D-Säulendiagramm

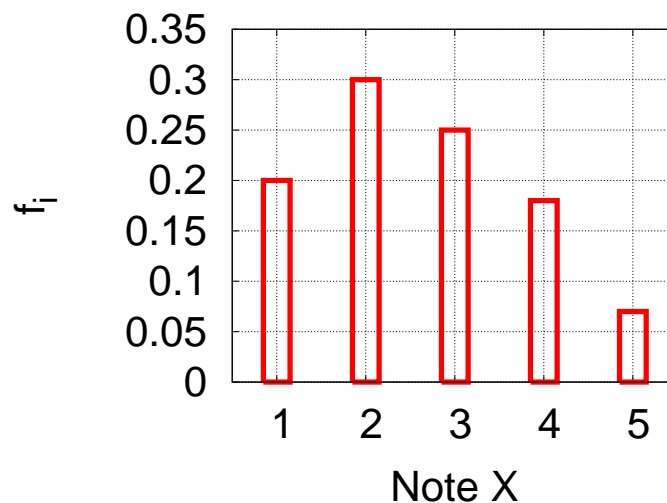


Würfeldiagramm

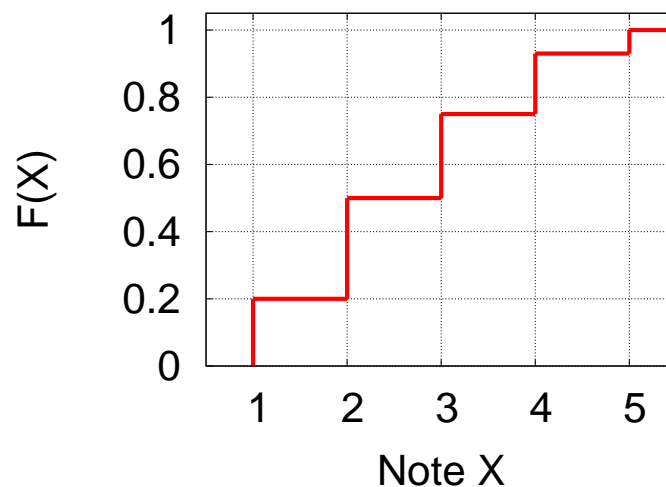


5.3(d) Darstellung nichtklassierter Daten

Sind die Daten mindestens ordinalskaliert, kann man das Stabdiagramm auch in einem "Koordinatensystem" darstellen:



Oder man plottet die relative Summenhäufigkeiten F_j bzw. die Verteilungsfunktion $F(x)$.



Diese letzte Darstellung ist *als einzige* auch für stetige Daten und/oder für sehr viele verschiedene Ausprägungen geeignet.

5.4: Klassierte Daten

Statistik ist für Politiker häufig das, was für Betrunkene die Laterne ist. Sie dient nicht der Erleuchtung, man klammert sich an ihr fest

Motivation:

- Hat man Daten mit sehr vielen unterschiedlichen Merkmalsausprägungen, wird die direkte Analyse der Daten nach Kap. 5.2, 5.3 unübersichtlich.
- Die direkte Analyse unklassierter Daten bietet keine sinnvolle Möglichkeit, die sehr anschauliche Häufigkeitsdichte bzw. Wahrscheinlichkeitsverteilung zu plotten.

Definition:

Klassierung bedeutet die Zusammenfassung sehr vieler **kardinalskalierter** Merkmalswerte zu relativ wenigen Klassen, d.h. Intervallen von Merkmalswerten: Die *i*-te **Klasse** eines Merkmals X ist durch das Intervall

$$x_k^u \leq X < x_k^o, \quad k = 1, \dots, K$$

gegeben. Dabei gilt

$x_{k+1}^u > x_k^u$: Die Klassen sind aufsteigend geordnet,

$x_{k+1}^u = x_k^o$: Die Intervalle stoßen aneinander.

5.4(b): Bemerkungen zur Klassierung

- Die erste bzw. letzte Klasse muss häufig als offene Klasse definiert werden: $x_1^u = -\infty$, $x_K^o = +\infty$. Bei der Bestimmung von Verteilungsfunktionen, Maßzahlen etc. sind dann sinnvolle *ad-hoc* Annahmen zu treffen!
- Die Wahl der Klassenanzahl K wird durch folgende Abwägung bestimmt.
 - Bei zu wenigen Klassen ergibt die Aggregation einen hohen Informationsverlust,
 - bei zu vielen Klassen hat man nicht gegenüber unklassifizierten Daten gewonnen.

Folgende Faustregel bewährt sich bei nicht zu vielen statistischen Einheiten ($n \leq 100$):

$$K \approx \sqrt{n}$$

Bei großem Datenumfang ($n > 100$) kann man z.B. DIN 55 302 heranziehen: ≥ 13 Klassen für $100 < n \leq 1000$, ≥ 16 Klassen zieht für $1000 < n \leq 10000$ usw.

- Man muss drei Sorten von Zählindices unterscheiden:
 1. Index für die Zählung der n *Elemente der Urliste*,
 2. Index für die Zählung der $m \leq n$ *verschiedenen Merkmalsausprägungen* in der Urliste,
 3. Index für die Zählung der K *Klassen*. Die Klassenzahl K muss nicht notwendigerweise kleiner als m oder n sein, dies ist aber sinnvoll (s.o.)

Alle drei Zählindices werden häufig in der Literatur (und auch in der Formelsammlung) mit denselben Buchstaben i und k bezeichnet! Was gemeint ist, muss aus dem Zusammenhang hervorgehen!

5.4(c): Klassenbreite, -mittel und -mitte

Die Definition dieser Größen ist für die Bestimmung der Häufigkeitsdichte und der Verteilungsfunktion sowie für die in Kap. 6 diskutierten Kenngrößen notwendig:

Klassenbreite:

$$\Delta_k = x_k^o - x_k^u$$

Das Klassenmittel

$$\bar{x}_k = \frac{1}{h_k} \sum_{i=1}^{h_k} (x_k)_i$$

ist das arithmetische Mittel der Elemente $(x_k)_k$ innerhalb der betrachteten Klasse i

Die Klassenmitte

$$x_k^* = \frac{x_k^o + x_k^u}{2}$$

ist das einfache Mittel aus der unteren und der oberen Klassengrenze.

Hinweis: Das Klassenmittel ist das “echte” Mittel der Merkmalswerte in der Klasse. Es ist genauer als die Klassenmitte, setzt aber die Kenntnis der Urliste voraus!

5.4(d): Verteilungsfunktion und Häufigkeitsdichte

Die

- **absoluten Klassenhäufigkeiten** h_k ,
- **relativen Klassenhäufigkeiten** $f_k = h_k/n$, sowie die
- **relativen Summenhäufigkeiten** $F_k = \sum_{i=1}^k f_i$

werden wie bei den nichtklassierten Daten ermittelt.

Unter der *Annahme einer Gleichverteilung der Merkmalswerte* innerhalb jeder der Klassen gibt es aber folgende neue Analysemöglichkeiten:

Absolute und relative Häufigkeitsdichte:

$$h_k^D = \frac{h_k}{\Delta_k}, \quad f_k^D = \frac{h_k}{n\Delta_k} = \frac{f_k}{\Delta_k}$$

Empirische Verteilungsfunktion:

$$F(x) = \begin{cases} 0 & \text{falls } x < x_1^u, \\ F_{k-1} + f_k \left(\frac{x - x_k^u}{\Delta_k} \right) & \text{falls } x_k^u \leq x < x_k^o \\ 1 & \text{falls } x > x_K^o \end{cases}$$

mit $F_0 = 0$, $k = 1, 2, \dots, K$,

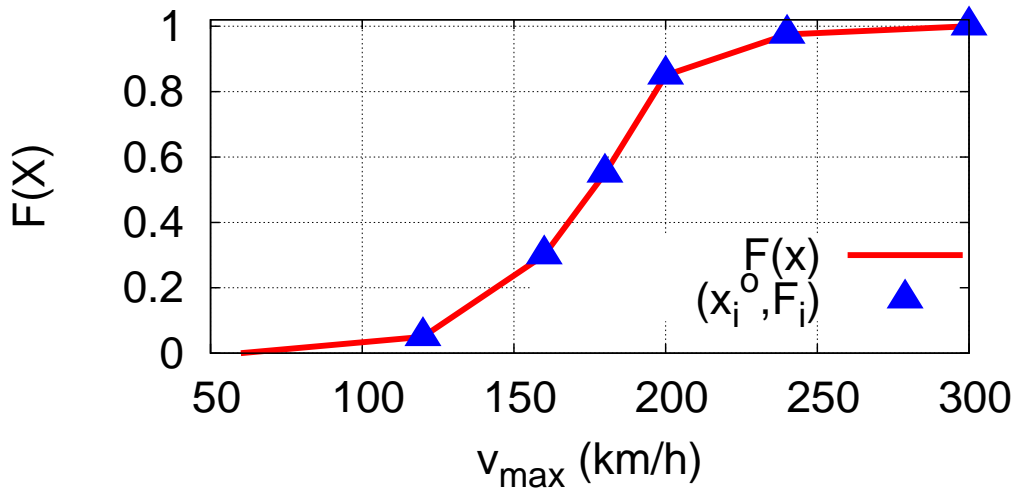
Empirische Dichtefunktion ("Histogramm")

$$f(x) = \frac{dF(x)}{dx} = \begin{cases} 0 & \text{falls } x < x_1^u, \\ f_k^D & \text{falls } x_k^u \leq x < x_k^o \\ 0 & \text{falls } x > x_K^o \end{cases}$$

mit $k = 1, 2, \dots, K$,

5.4(e): Grafische Darstellung klassierter Daten

Verteilungsfunktion



Relative Häufigkeitsdichte

