Vorlesung Statistik A
Explorative Datenanalyse

Florian Wickelmaier

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN



Wintersemester 2026/2027

Häufigkeitsverteilungen von diskreten Merkmalen

Wir betrachten ein diskretes Merkmal A mit seinen Merkmalsausprägungen

$$A = \{a_1, \ldots, a_j, \ldots, a_m\}$$

Beispiel: Noten bei Klassenarbeit

- Urliste für $n = 25$

$$\mathbf{x} = (3, 5, 4, 3, 2, 3, 4, 6, 1, 2, 3, 3, 4, 5, 2, 1, 3, 4, 2, 4, 3, 1, 2, 3, 4)$$

Beispiel: Noten bei Klassenarbeit

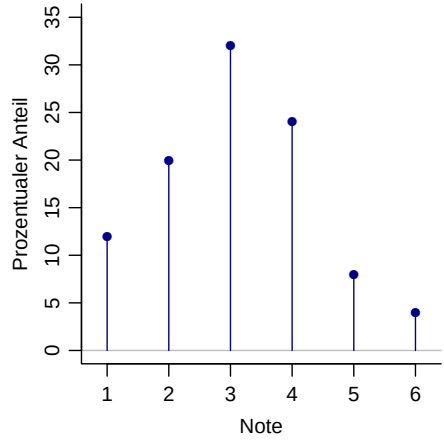
Urliste

$$\mathbf{x} = (3, 5, 4, 3, 2, 3, 4, 6, 1, 2, 3, 3, 4, 5, 2, 1, 3, 4, 2, 4, 3, 1, 2, 3, 4)$$

Häufigkeitstabelle

A	Strichliste	h_j	f_j	$100\% \cdot f_j$
1	|||	3	0.12	12
2	|||||	5	0.20	20
3	||||| |||	8	0.32	32
4	||||| |	6	0.24	24
5	||	2	0.08	8
6	|	1	0.04	4
$\sum$		25	1.000	100.0

Stabdiagramm



Empirische Verteilungsfunktion

Bei vielen Problemen möchte man wissen, wieviele der Beobachtungen in der Urliste einen bestimmten Grenzwert x nicht überschreiten

- Beachte: Diese Fragestellung ist nur für Merkmale mit mindestens Ordinalskalenniveau sinnvoll

Gesucht ist also die Anzahl von Beobachtungen $H(x)$, die kleiner oder gleich dem fest vorgegebenen Zahlenwert x sind



Empirische Verteilungsfunktion

Absolute kumulierte Häufigkeit

$$H(x) = \sum_{j:\ a_j \leq x} h(a_j)$$

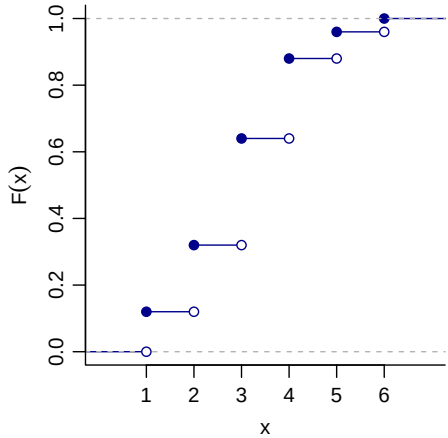
Relative kumulierte Häufigkeit

$$F(x) = \sum_{j:\ a_j \leq x} f(a_j) = \frac{H(x)}{n}$$

Empirische Verteilungsfunktion

Beispiel: Noten bei Klassenarbeit

a_j	$h(a_j)$	$H(a_j)$	$f(a_j)$	$F(a_j)$
1	3	3	0.12	0.12
2	5	8	0.20	0.32
3	8	16	0.32	0.64
4	6	22	0.24	0.88
5	2	24	0.08	0.96
6	1	25	0.04	1.00



Empirische Verteilungsfunktion

Eigenschaften von $F(x)$

- $F(x) = 0$ für $x < a_1$
- $F(x) = 1$ für $x \geq a_m$
- $F(x)$ ist eine monoton nicht-fallende Funktion:
  Für $x < x'$ ist $F(x) \leq F(x')$

Empirische Verteilungsfunktion in R

Beispiel: Noten bei Klassenarbeit

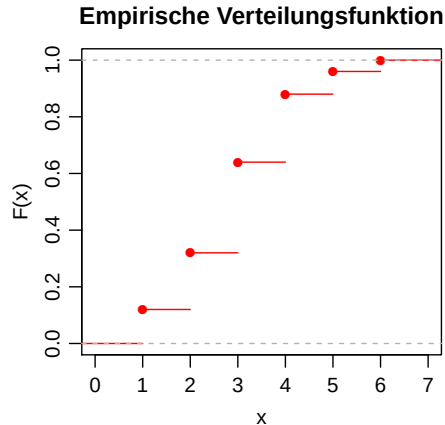
- Für einen gegebenen Datenvektor (Urliste) berechnet die Funktion `ecdf()` die empirische kumulative Verteilungsfunktion
- Das Ergebnis kann unmittelbar grafisch dargestellt werden

```
x <- c(3,5,4,3,2,3,4,6,1,2,3,3,4,5,2,1,3,4,2,4,3,1,2,3,4)
F <- ecdf(x)  # Rueckgabewert ist eine Funktion
plot(F, col.points = "red", col.hor = "red", pch = 16,
     ylab = "F(x)", main = "Empirische Verteilungsfunktion")
```

- Argumente der Plotmethode
  - `col.points` und `col.hor` setzen die Farbe der Punkte und horizontalen Linien
  - `pch = 16` ersetzt das Standardplotsymbol durch einen ausgefüllten Kreis

Empirische Verteilungsfunktion in R

Beispiel: Noten bei Klassenarbeit



Häufigkeitsverteilungen bei (quasi-)stetigen Merkmalen

Bei einem stetigen Merkmal sind im Allgemeinen die Beobachtungswerte in der Urliste allesamt verschieden

- Auch nach Runden der Werte werden diese nur selten übereinstimmen
- Die Häufigkeitsverteilungen werden in diesem Fall sehr unübersichtlich

Bei diskreten Merkmalen mit sehr vielen Merkmalsausprägungen tritt dasselbe Problem auf

- Derartige Merkmale werden als *quasi-stetig* bezeichnet und wie stetige Merkmale behandelt

In diesen Fällen nimmt man eine *Klasseneinteilung* (Gruppierung, Kategorisierung) der Merkmalsausprägungen vor

Klasseneinteilung

Beispiel: Körpergewicht von 100 männlichen Studenten

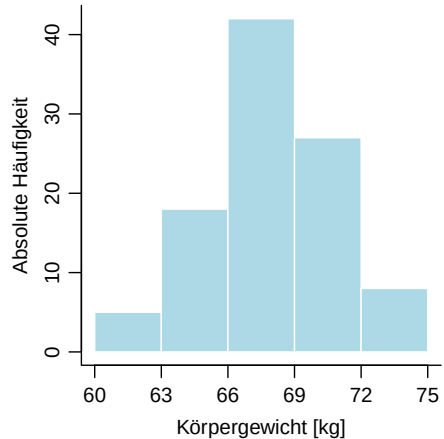
Gewicht (kg)	Anzahl der Studenten
$(60, 63]$	5
$(63, 66]$	18
$(66, 69]$	42
$(69, 72]$	27
$(72, 75]$	8

Dabei bezeichnet $(a, b] = \{x \mid a < x \leq b\}$ ein *halboffenes* Intervall (genauer, ein *linksoffenes* Intervall)

Häufigkeitstabelle und Histogramm

Beispiel: Körpergewicht von 100 männlichen Studenten

Klasse	Klassengrenzen	h_j	H_j	f_j	F_j
K_1	$(60, 63]$	5	5	0.05	0.05
K_2	$(63, 66]$	18	23	0.18	0.23
K_3	$(66, 69]$	42	65	0.42	0.65
K_4	$(69, 72]$	27	92	0.27	0.92
K_5	$(72, 75]$	8	100	0.08	1.00



Häufigkeitsverteilungen bei (quasi-)stetigen Merkmalen

Die Klassen

$$K_j = (u_{j-1}, u_j] = \{x \mid u_{j-1} < x \leq u_j\}$$

für $j = 1, \ldots, m$ sind durch die Grenzen $u_0, u_1, \ldots, u_m$ festgelegt

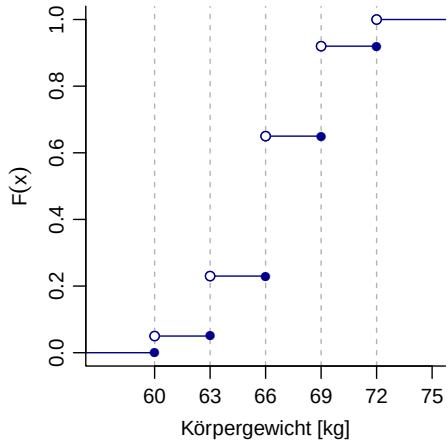
Für die Verteilungsfunktion gilt

$$F(u_0) = 0$$

$$F(u_k) = \sum_{j=1}^{k} f_j$$

Treppenfunktion

Beispiel: Körpergewicht von 100 männlichen Studenten



Treppenfunktion

Die Treppenfunktion ist als Näherung der eigentlichen Verteilungsfunktion des unterliegenden (quasi-)stetigen Merkmals zu sehen

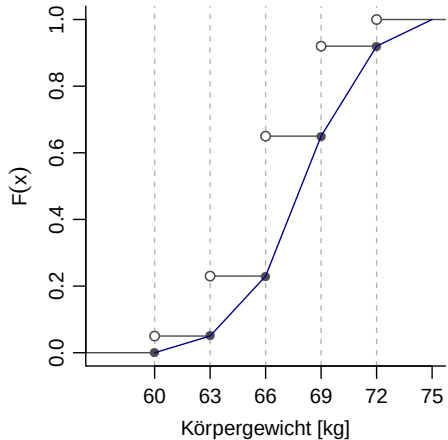
Sie wird umso "glatter" aussehen, je kleiner die Klassenbreiten gewählt werden

Alternativ zur Treppenfunktion kann man auch das *kumulative Häufigkeitspolygon* betrachten

- Hierbei wird F_j gegen die obere Klassengrenze u_j der Klasse K_j abgetragen

Häufigkeitspolygon

Beispiel: Körpergewicht von 100 männlichen Studenten



Bestimmung der optimalen Klassenbreite

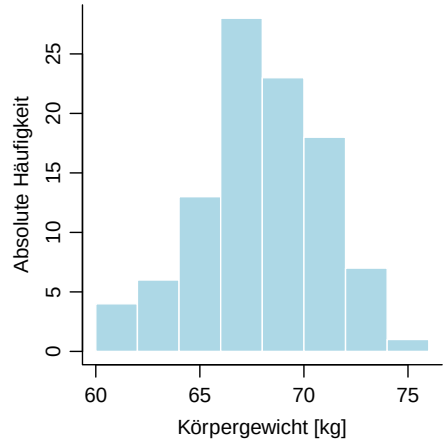
Faustregel (nach Sturges, 1926)

- Anzahl der Klassen

$$m = 1 + \log_2(n) = 1 + 3.32 \cdot \log_{10}(n)$$

- Für das Beispiel mit $n = 100$ erhält man $m = 7.64 \approx 8$
- Die Klassenbreite berechnet sich dann durch

$$\frac{x_{max} - x_{min}}{m}$$



Ungleiche Klassenbreiten

Bei ungleichen Klassenbreiten (etwa zur Vermeidung leerer Klassenbesetzungen) werden flächenproportionale Histogramme eingesetzt

Dabei repräsentieren die Flächen der Histogramm-Rechtecke die jeweilige Klassenhäufigkeit

Bei gegebener Klassenbreite $b_j = u_j - u_{j-1}$ und relativer Klassenhäufigkeit f_j berechnet sich die Rechteckhöhe y_j ("Dichte") für die Klasse K_j, $j = 1, \ldots, m$, durch

$$y_j = \frac{f_j}{b_j} = \frac{\text{relative Klassenhäufigkeit}}{\text{Klassenbreite}}$$

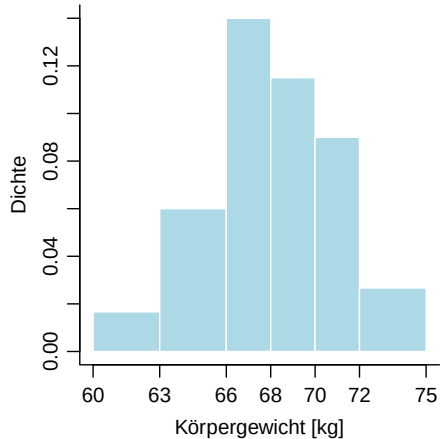
Ungleiche Klassenbreiten

Beispiel: Körpergewicht von 100 männlichen Studenten

Klasse	Klassengrenzen	h_j	H_j	f_j	F_j	b_j	y_j
K_1	$(60, 63]$	5	5	0.05	0.05	3	0.017
K_2	$(63, 66]$	18	23	0.18	0.23	3	0.060
K_3	$(66, 68]$	28	51	0.28	0.51	2	0.140
K_4	$(68, 70]$	23	74	0.23	0.74	2	0.115
K_5	$(70, 72]$	18	92	0.18	0.92	2	0.090
K_6	$(72, 75]$	8	100	0.08	1.00	3	0.027

Ungleiche Klassenbreiten

Beispiel: Körpergewicht (flächenproportionales Histogramm)



Häufigkeitsverteilungen bei (quasi-)stetigen Merkmalen

Histogramme in R mit Funktion `hist()`

Beispiel: Körpergewicht von 100 männlichen Studenten

```
## Urliste
x <- c(61.1,62.0,62.2,61.8,60.2,63.4,63.5,64.1,65.0,64.2,
       67.5,65.4,64.7,64.4,65.1,65.2,68.2,63.8,63.6,65.3,
       66.9,63.4,65.0,67.2,67.2,67.2,68.4,66.6,67.4,67.6,
       66.0,68.7,66.8,66.1,66.3,66.5,68.9,67.9,69.4,67.7,
       66.4,68.4,67.7,67.2,66.7,66.0,67.8,67.8,67.7,68.2,
       68.8,66.5,68.8,66.8,66.1,67.1,68.7,68.8,68.9,66.8,
       68.3,68.3,67.9,67.3,66.0,70.9,70.5,69.0,69.1,69.7,
       69.6,70.9,70.1,70.1,69.3,71.1,70.2,71.7,69.3,71.9,
       71.1,70.9,71.6,71.0,70.6,69.5,69.1,71.3,70.2,71.1,
       69.5,71.7,73.8,72.3,72.2,73.0,72.4,74.9,72.5,72.3)
```

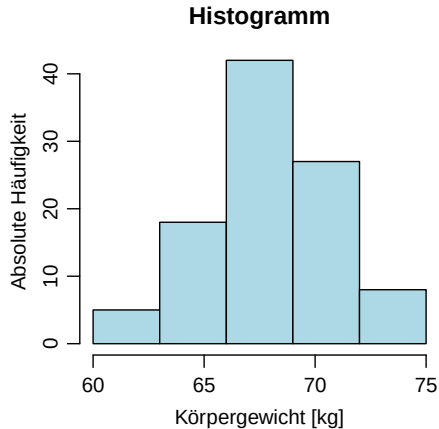
Histogramme in R mit Funktion `hist()`

```
hist(x, breaks = c(60,63,66,69,72,75), freq = TRUE, right = TRUE,
     col = "lightblue", main = "Histogramm",
     xlab = "Körpergewicht [kg]",
     ylab = "Absolute Häufigkeit")
```

- `breaks` gibt die Kategoriengrenzen an
- Mit `freq = TRUE` werden absolute Häufigkeiten ausgegeben, mit `freq = FALSE` die Dichte (Rechteckhöhe bei flächenproportionalem Diagramm)
- Mit `right = TRUE` sind die Kategorien nach links halboffene Intervalle
- `col` definiert die Farbe des Histogramms
- Mit `main`, `xlab`, `ylab` spezifiziert man den Titel des Diagramms sowie die Beschriftung der x- und y-Achse

Histogramme in R mit Funktion `hist()`

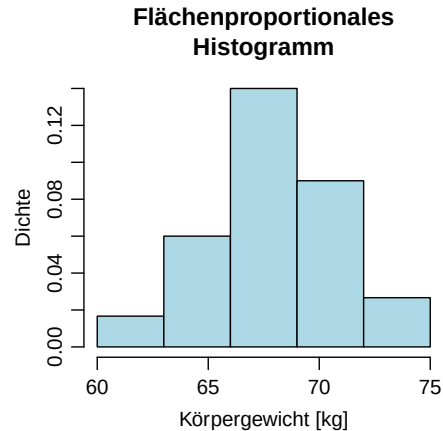
Beispiel: Körpergewicht von 100 männlichen Studenten



Histogramme in R mit Funktion `hist()`

Flächenproportionales Histogramm

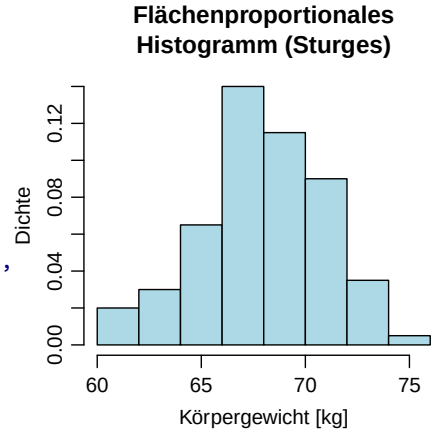
```
hist(x, breaks = c(60,63,66,69,72,75),
     freq = FALSE, right = TRUE,
     col = "lightblue",
     main = "Flächenproportionales\nHistogramm",
     xlab = "Körpergewicht [kg]", ylab = "Dichte")
```



Histogramme in R mit Funktion `hist()`

Kategoriengrenzen nach Sturges

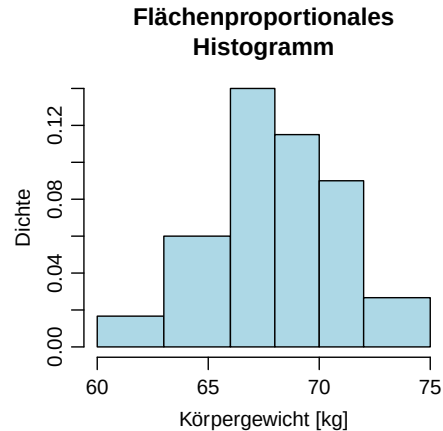
```
hist(x, breaks = "Sturges",
     freq = FALSE, right = TRUE,
     col = "lightblue",
     main =
      "Flächenproportionales\nHistogramm (Sturges)",
     xlab = "Körpergewicht [kg]", ylab = "Dichte")
```



Histogramme in R mit Funktion `hist()`

Ungleiche Kategorienbreite

```
hist(x, breaks = c(60,63,66,68,70,72,75),
     freq = FALSE, right = TRUE,
     col = "lightblue",
     main = "Flächenproportionales\nHistogramm",
     xlab = "Körpergewicht [kg]", ylab = "Dichte")
```



Histogramme in R mit Funktion `hist()`

Rückgabewert

```
hist(x, breaks = c(60,63,66,69,72,75), plot = FALSE)
```

Mit `plot = FALSE` wird kein Diagramm gezeichnet (Voreinstellung ist `plot = TRUE`), sondern eine Liste ausgegeben

```
$breaks
[1] 60 63 66 69 72 75
$counts
[1]  5 18 42 27  8
$density
[1] 0.01666666 0.06000000 0.14000000 0.09000000 0.02666667
$mids
[1] 61.5 64.5 67.5 70.5 73.5
```

Histogramme in R mit Funktion `hist()`

Rückgabewert

```
h <- hist(x, breaks = "Sturges", plot = TRUE)
```

Mit dieser Zuweisung wird ein Diagramm gezeichnet und die zuvor genannten Informationen sind folgendermaßen abrufbar

```
h$breaks
[1] 60 62 64 66 68 70 72 74 76

h$counts
[1]  4  6 13 28 23 18  7  1

round(h$density, digits = 3)
[1] 0.020 0.030 0.065 0.140 0.115 0.090 0.035 0.005
```

Histogramme in R mit Funktion `hist()`

Histogramm in R mit angepasster Achsenbeschriftung

```
hist(x, breaks = c(60,63,66,69,72,75),
     freq = TRUE, right = TRUE, col = "lightblue",
     main = "Histogramm", xlab = "Körpergewicht [kg]",
     ylab = "Absolute Häufigkeit", axes = FALSE)
axis(1, at = seq(60, 75, 3))
axis(2)
```

- Funktion `hist(..., axes = FALSE)` stellt das Histogramm ohne Achsen dar
- Der Aufruf `axis(1, at = seq(60, 75, 3))` zeichnet auf der x-Achse die Ticks von 60 bis 75 in einem Abstand von 3 ein
- Mit `axis(2)` wird die standardmäßig generierte y-Achse hinzugefügt

Histogramme in R mit Funktion `hist()`

Beispiel: Körpergewicht von 100 männlichen Studenten

