Einführende Statistik

Tabellarische und graphische Aufbereitung von Daten

Marcus Hudec

Absolute Häufigkeiten diskreter Merkmale

Wir betrachten ein Merkmal X, das entweder

ein diskretes quantitatives Merkmal mit k unterschiedlichen Realisationsmöglichkeiten x_i mit i=1,...,k

ist, oder

ein qualitatives Merkmal, das entweder eine Nominal- oder eine Ordinalskala aufweist

- Die Anzahl des Vorkommens von x_i in einer Population heißt absolute Häufigkeit:

$$n(X=x_i) \text{ bzw. kurz } n_i \qquad n() \ldots \text{Zählfunktion}$$

Beispiel: Personalbeurteilung (Schlittgen p.13)

- n=120 Beurteilungsbögen
- Merkmal: "Produktives Denken"
- Ausprägungen x_i (k=10):
  0 ... sehr gering
  bis
  9 ... sehr gut ausgeprägt
- Skalenniveau: Ordinal
- Urliste:
  6 7 7 7 5 5 ...

  ... 3 5 6 5

Häufigkeitstabelle (absolute Häufigkeiten)

i	x_i	n_i
1	0	0
2	1	0
3	2	0
4	3	7
5	4	12
6	5	38
7	6	29
8	7	27
9	8	6
10	9	1
Gesamt		120

i … laufender Index

x_i …Ausprägungen

n_i … Häufigkeiten

Häufigkeitstabellen mit R

```
> # ================================================================================
> # Start der Session
> # ================================================================================
>
> # Wechsel auf das Directory mit Daten
> setwd("C:\\work\\data\\")
> # Analyse von Daten
> prod.denken <- read.csv(file="Produktiv.csv")
> daten <- as.vector(as.matrix(prod.denken))
> table(daten)
daten
 3  4  5  6  7  8  9
 7 12 38 29 27  6  1
> table.fix(daten, von=0, bis=9)
 0  1  2  3  4  5  6  7  8  9
 0  0  0  7 12 38 29 27  6  1
> # Relative Häufigkeiten
> table.fix(daten, von=0, bis=9)/length(daten)
          0           1           2           3           4           5
0.000000000 0.000000000 0.000000000 0.058333333 0.100000000 0.316666667
          6           7           8           9
0.241666667 0.225000000 0.050000000 0.008333333
>
```

Relative Häufigkeiten diskreter Merkmale

- Die relative Häufigkeit h_i von einer Merkmalsausprägung x_i ist wie folgt definiert:

  die absolute Häufigkeit n_i der Merkmalsausprägung x_i dividiert durch die Gesamtzahl der Merkmalsträger n

$$h(X = x_i) \equiv \frac{n(X = x_i)}{n} = \frac{n_i}{n} \equiv h_i$$

$$n = \sum_{i=1}^{k} n_i \qquad \sum_{i=1}^{k} h_i = 1$$

Häufigkeitstabelle (absolute & relative Häufigkeiten)

i	x_i	n_i	h_i	h_i in %
1	0	0	0,00	0,00%
2	1	0	0,00	0,00%
3	2	0	0,00	0,00%
4	3	7	0,06	5,83%
5	4	12	0,10	10,00%
6	5	38	0,32	31,67%
7	6	29	0,24	24,17%
8	7	27	0,23	22,50%
9	8	6	0,05	5,00%
10	9	1	0,01	0,83%
Gesamt		120	1	100,00%

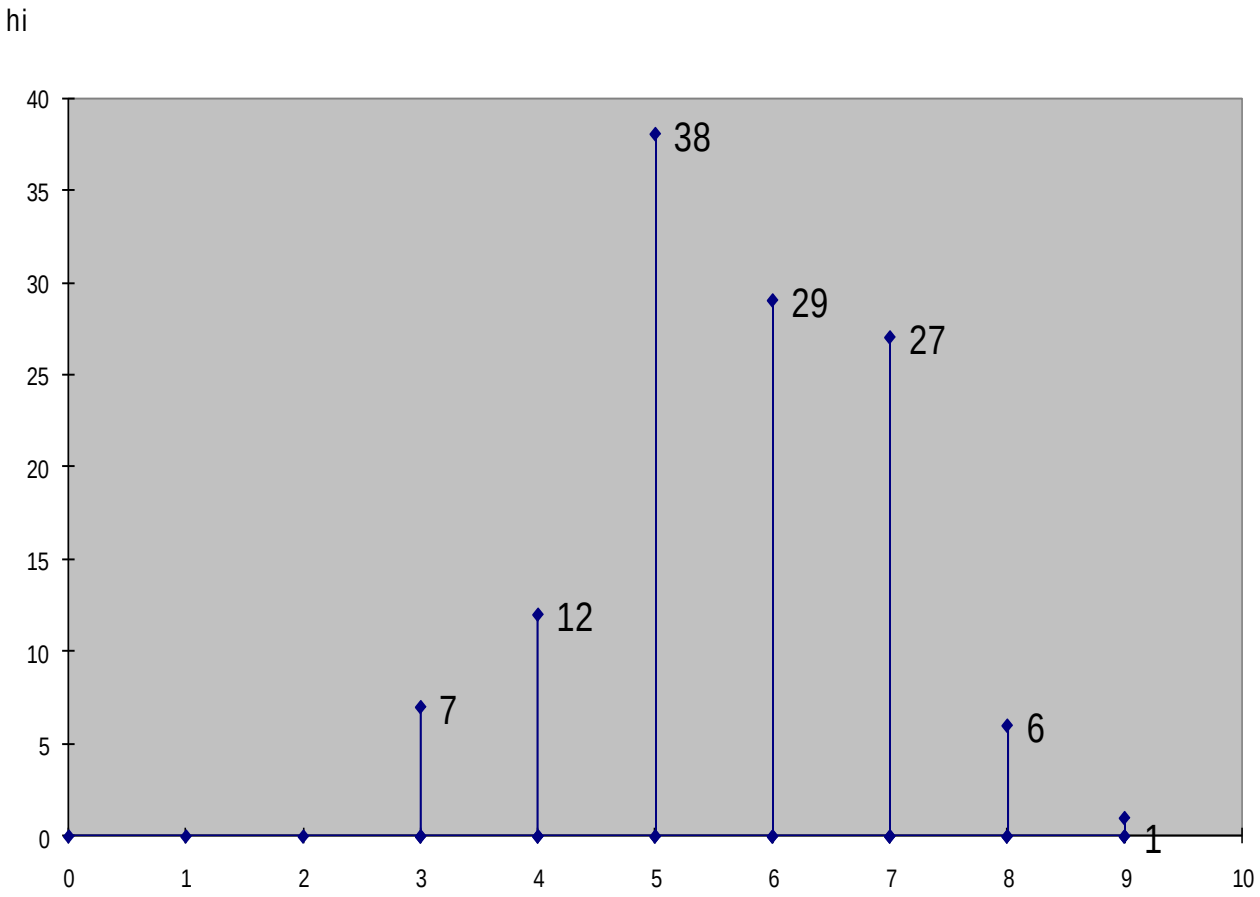
In der Praxis werden die relativen Häufigkeiten oft mit 100 multipliziert und als Prozentwerte dargestellt

Erzeugen der Tabelle mit R

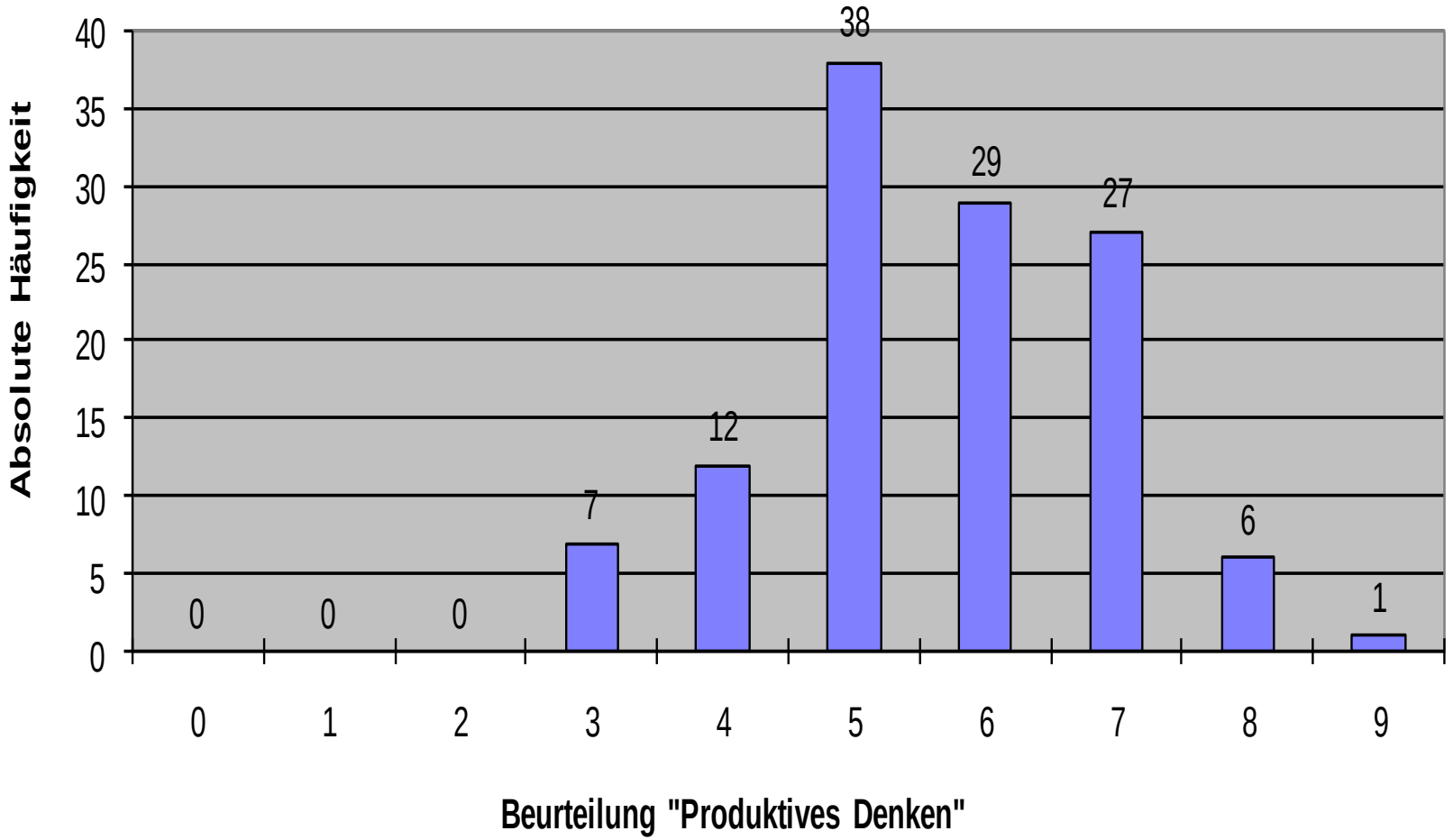
```
ni <- table.fix(daten, von=0, bis=9)
Tabelle <- cbind(ni, ni/length(daten))
Tabelle
rownames(Tabelle) <- paste("X =", 0:9)
colnames(Tabelle) <- c("abs. Häuf.", "rel. Häuf")
Tabelle
```

```
> Tabelle
      abs. Häuf.   rel. Häuf
X = 0          0 0.000000000
X = 1          0 0.000000000
X = 2          0 0.000000000
X = 3          7 0.058333333
X = 4         12 0.100000000
X = 5         38 0.316666667
X = 6         29 0.241666667
X = 7         27 0.225000000
X = 8          6 0.050000000
X = 9          1 0.008333333
```

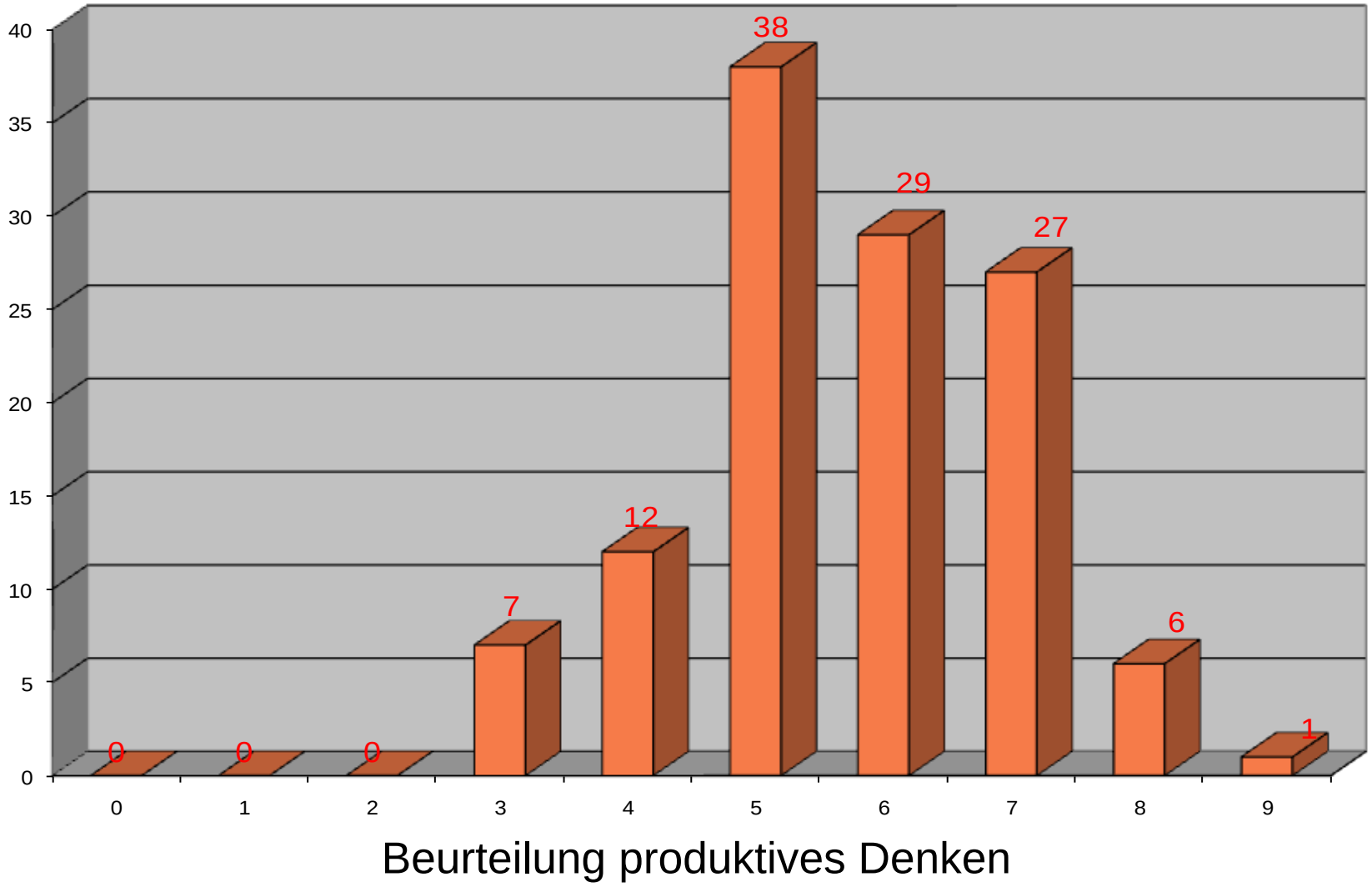
Stabdiagramm: Produktives Denken mit Excel



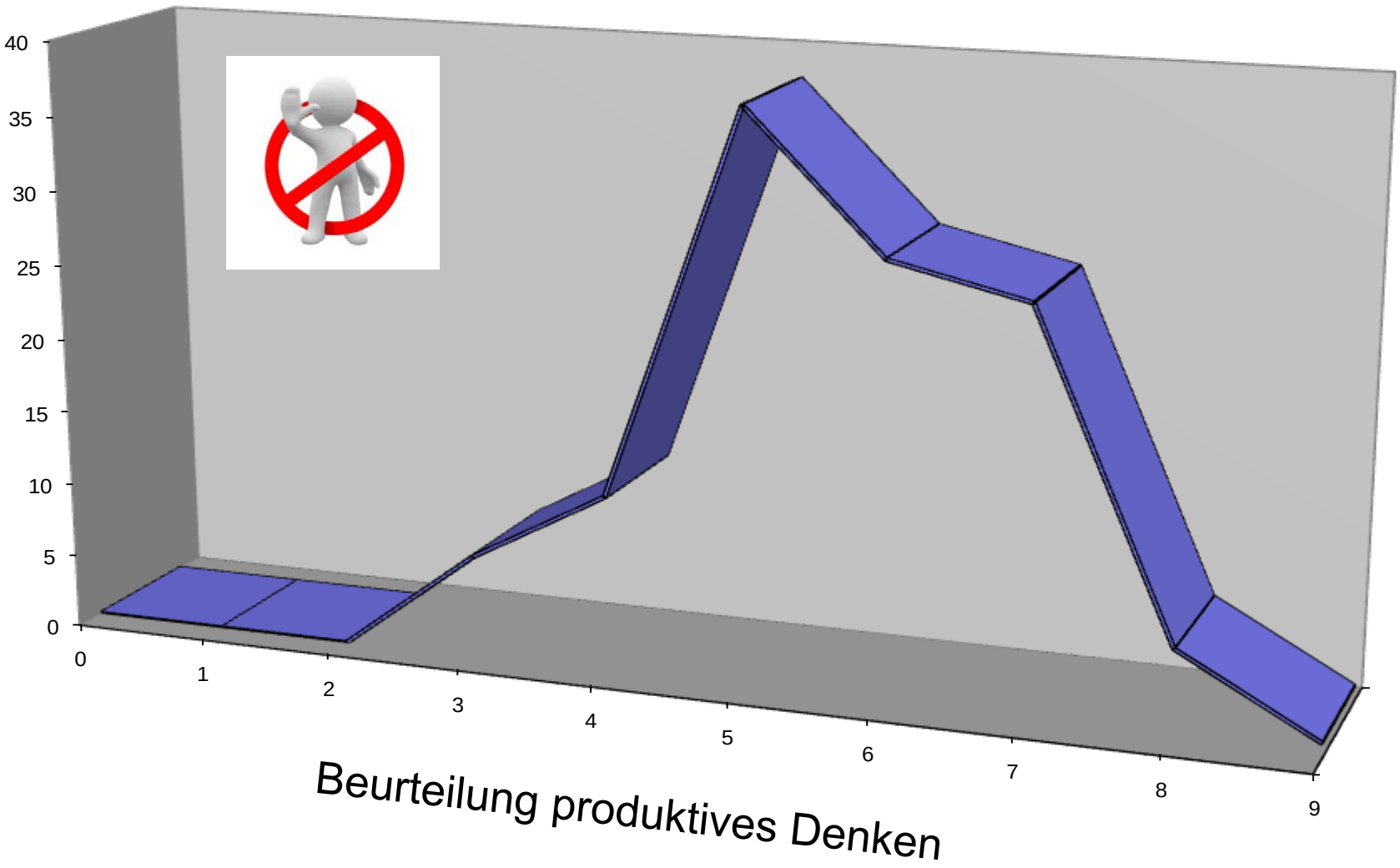
Säulendiagramm: Produktives Denken



3-dimensionales Säulendiagramm

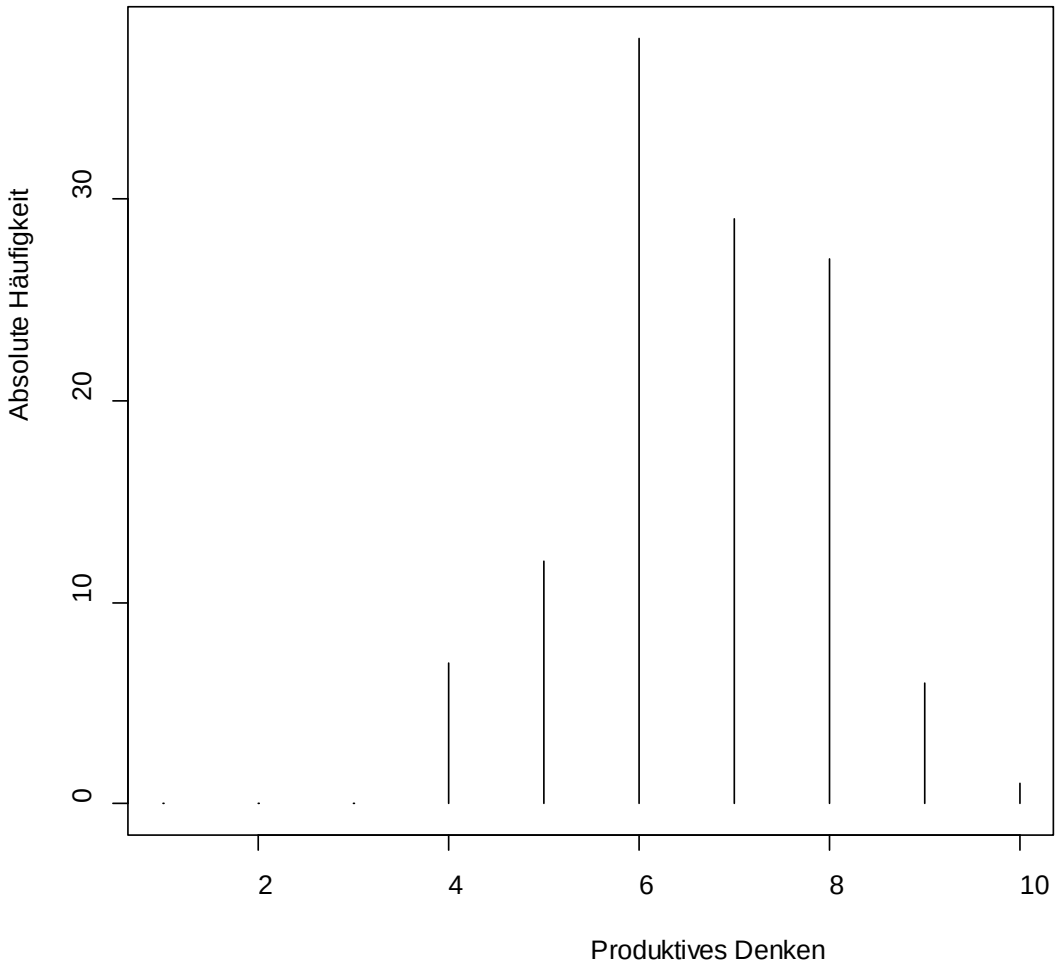


Kontraproduktives Design

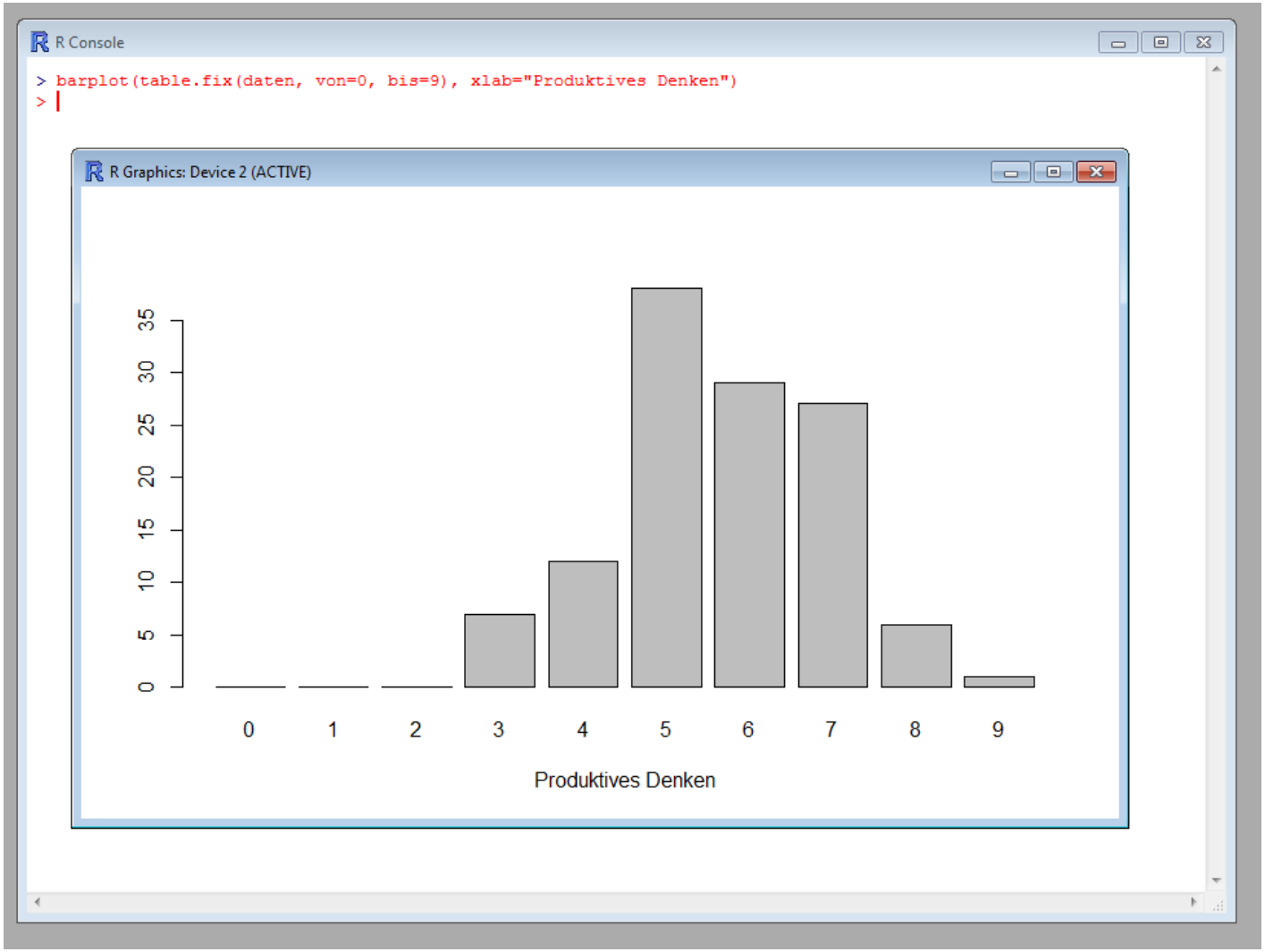


Stabdiagramm mit R

```
plot(table.fix(daten, von=0, bis=9), type="h",
      xlab="Produktives Denken", ylab = "Absolute Häufigkeit")
```



Säulendiagramm mit R



4 Leitregeln für statistische Graphiken

- Eine statistische Graphik sollte möglichst selbsterklärend sein.
  - Möglichst exakte Angabe der Datenquelle bzw. der Grundgesamtheit.
- Es sollte möglich sein, auf die zugrundeliegenden numerischen Daten rückschließen zu können
  - Gitterlinien, Wertangaben
- Die erste optische Wahrnehmung muss die tatsächlichen Größenordnungen korrekt widerspiegeln.
  - Menschen nehmen Flächen wahr ➔ Flächen müssen die Quantität reflektieren
- Die Graphik soll optisch attraktiv sein, aber eine klare Botschaft vermitteln.
  - Erwecken des Interesses des Betrachters

Vier zentrale Prinzipien statistischer Graphiken

- **Selbsterklärend**
  (Qualitative Information)
- **Numerische Transparenz**
  (Quantitative Information)
- **Graphische Integrität**
  (Korrektheit der Information)
- **Optische Attraktivität**
  (Anziehungskraft)

Daten aus Kühnel & Krebs p.44

Frage nach der Einschätzung der Wirtschaftlage in Deutschland (1996)

Ausprägung	Code	Häufigkeit
sehr gut	1	30
gut	2	435
teils/teils	3	1710
schlecht	4	1087
sehr schlecht	5	232
keine Antwort	-99	24

Daten aus Kühnel & Krebs p.44 - mit Excel aufbereitet

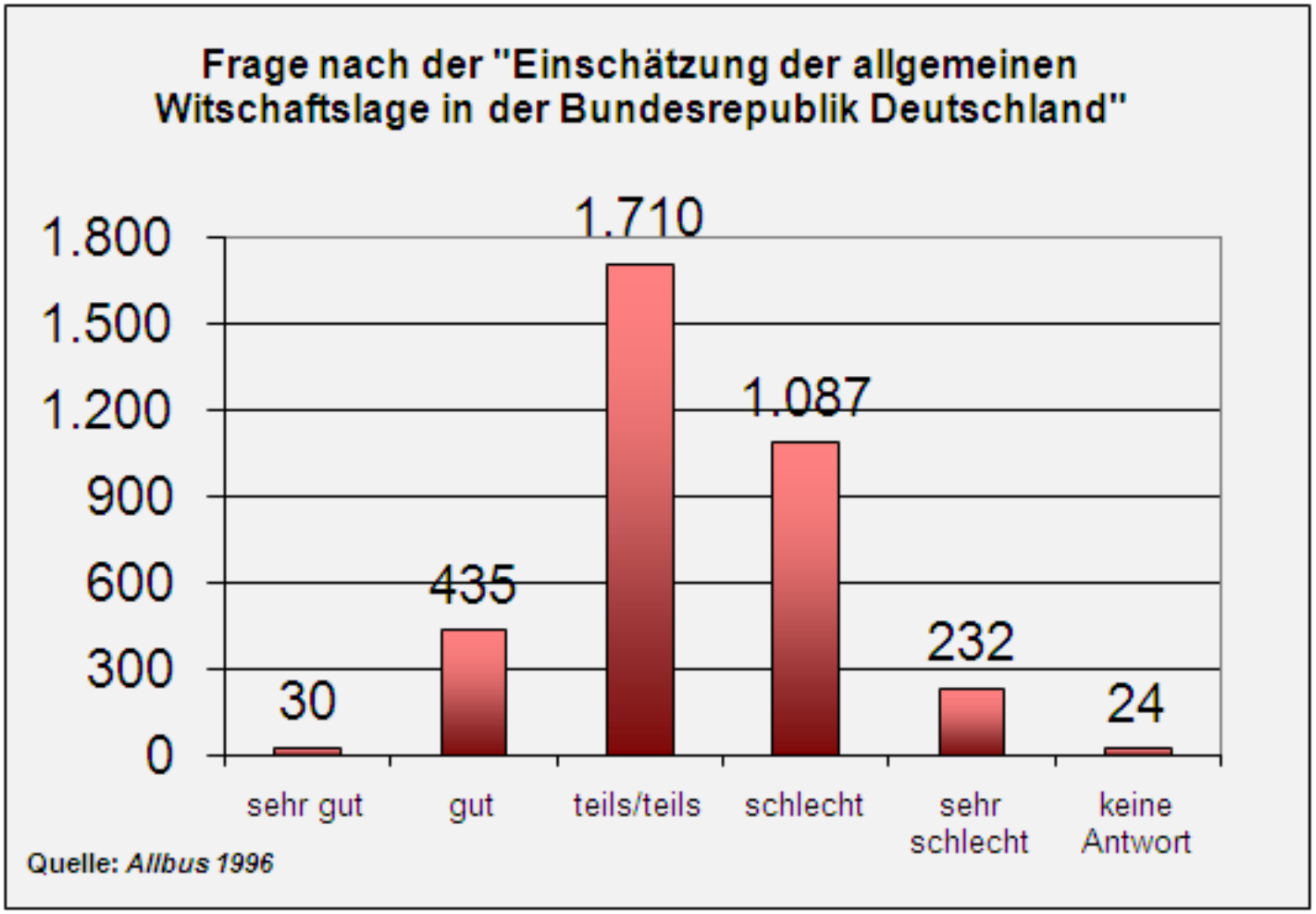
Frage nach der "Einschätzung der allgemeinen Witschaftslage in der Bundesrepublik Deutschland"

Ausprägung	Code	Häufigkeit	Prozente	Gültige Prozente	Kumulierte Prozente
sehr gut	1	30	0,9%	0,9%	0,9%
gut	2	435	12,4%	12,4%	13,3%
teils/teils	3	1.710	48,6%	48,9%	62,2%
schlecht	4	1.087	30,9%	31,1%	93,4%
sehr schlecht	5	232	6,6%	6,6%	100,0%
keine Antwort	-99	24	0,7%	Missing	
Total		3.518	100,0%	100,0%	
Gültige Fälle:	3.494	Fehlende Fälle:	24		

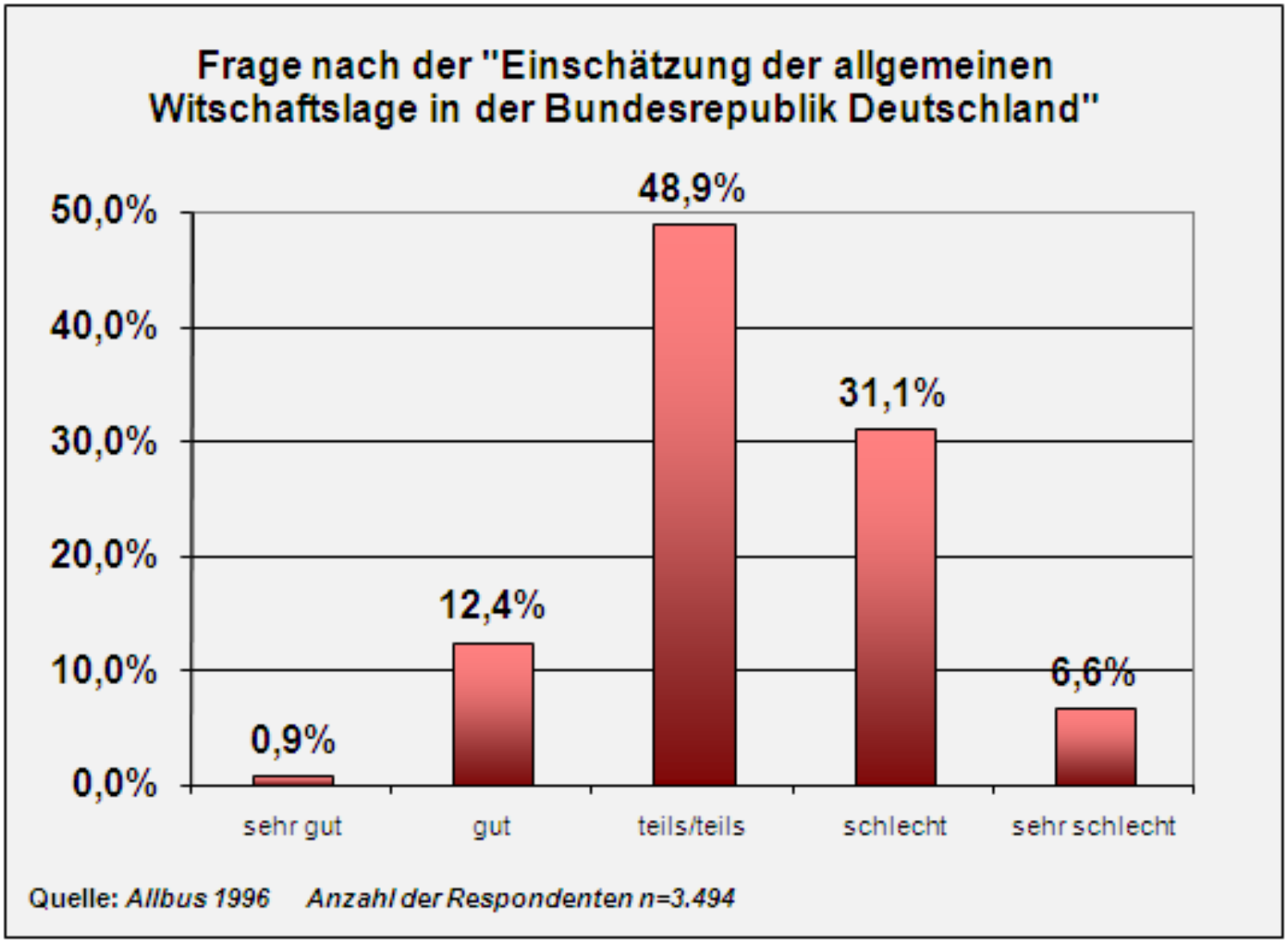
Quelle: *Allbus 1996*

Die gültigen Prozentwerte beziehen die absolute Häufigkeit gültiger Antworten auf die Gesamtzahl aller gültigen Antworten

Säulendiagramm mit absoluten Häufigkeiten



Säulendiagramm mit relativen Häufigkeiten



Beispiel zum Umgang mit fehlenden Werten

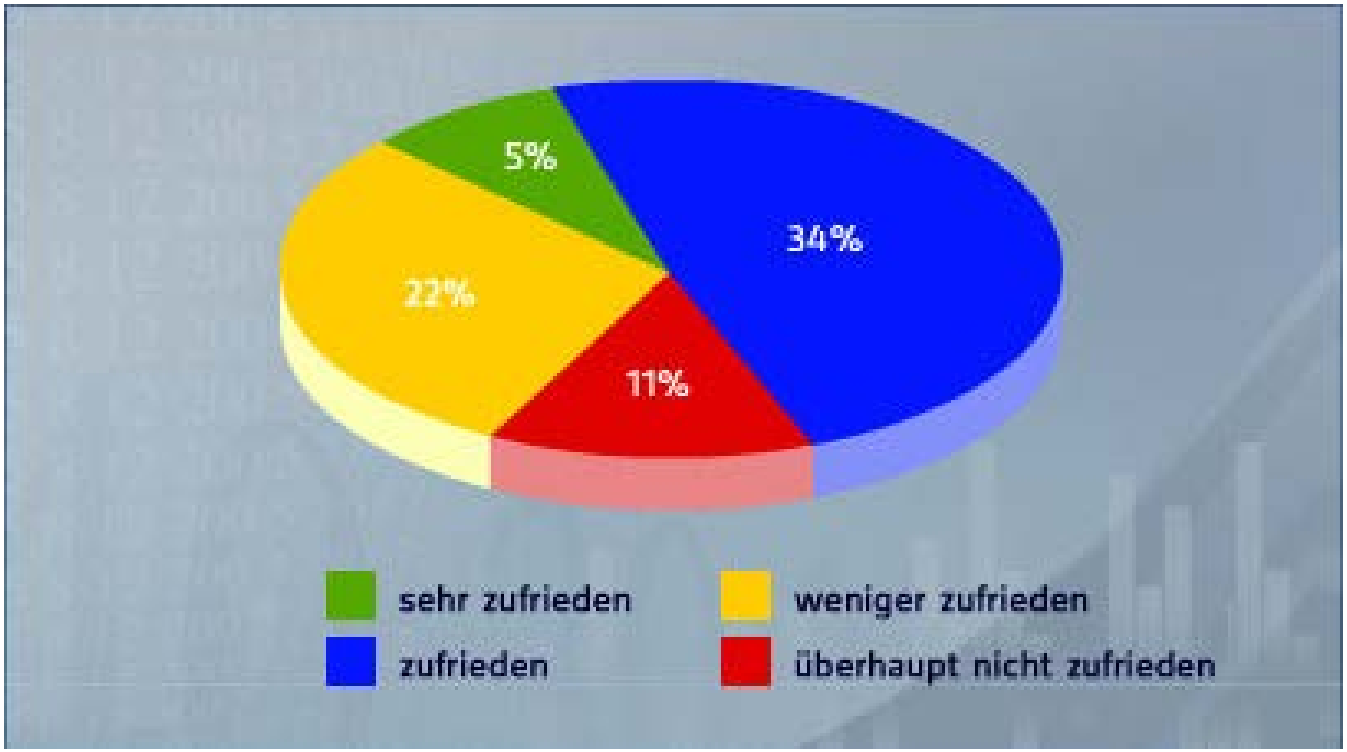
	Absolute Häufigkeit	relative Häufigkeit in Prozent	gültige Prozente	fiktive Anzahl	Aufteilung der Nicht-Antworter	relative Aufteilung der Nichtantworter
Partei A	3.000	30%	37,50%	3.750	750	37,50%
Partei B	2.500	25%	31,25%	3.125	625	31,25%
Partei C	1.500	15%	18,75%	1.875	375	18,75%
Partei D	1.000	10%	12,50%	1.250	250	12,50%
Keine Antwort	2.000	20%				
	10.000	100%	100%		2.000	
# gültige Antworten	8.000					

Wir erkennen, dass bei der Berechnung gültiger Prozent implizit unterstellt wird, dass sich die Non-Responder gleich wie die Responder verhalten.
Bei Kenntnis eines abweichenden Verhaltens der Non-Responder mag eine alternative Aufteilung der Antwortausfälle sinnvoll sein!

Umgang mit fehlenden Antworten

3 Monate vor der Fußball-WM 2006 war die Kampagne der Bild-Zeitung gegen den Bundestrainer Jürgen Klinsmann in vollem Gange. Auf der Online-Ausgabe "bild.de" konnte man bei der Sonntags-Frage zur WM am 12.3.2006 nachlesen:

"Nur fünf Prozent sind sehr zufrieden mit Klinsmann“



Korrekte Darstellung

- 5% aller Befragten ~ 6,9% aller Antwortenden

Antwort	Prozent	Gültige Prozent
sehr zufrieden	5%	6,9%
zufrieden	34%	47,2%
weniger zufrieden	11%	15,3%
überhaupt nicht zufrieden	22%	30,6%
keine Antwort	28%	---
	100%	100,0%

- Beachte: Keine Angabe über die Zahl der Respondenten
- Repräsentativität ?
  Unzufriedene äußern sich z.B. wesentlich häufiger in Web-Foren
- Manipulation: 54,1% der Antworter sind mit Klinsmann zufrieden oder sogar sehr zufrieden

Imputation fehlender Werte

- **Imputationsmethoden** haben das Ziel fehlende Werte möglichst sinnvoll zu ergänzen
- Grundgedanke ist dabei das Ausnützen von Abhängigkeiten zwischen Merkmalen
- Unterschiedliche Ergebnisse bei politischen Meinungsumfragen unterscheiden sich häufig nur in der methodischen Behandlung fehlender Werte:
  Wie werden die Unschlüssigen bzw. die Antwortverweigerer aufgeteilt?

Absolute oder relative Häufigkeiten?

- Für Vergleichszwecke eignen sich relative Häufigkeiten natürlich besser
- Bei Stichproben interessieren meist die Anteile (~relativen Häufigkeiten), aber der Umfang der Stichprobe muss unbedingt kommuniziert werden, um die Relevanz der Ergebnisse beurteilen zu können.
- Absolute Häufigkeiten kommunizieren stärker die Betroffenheit:
  - Im Jahresdurchschnitt gab es in Österreich im Jahr 2013 laut AMS 287.200 als arbeitslos vorgemerkte Personen
  - Im Jahresdurchschnitt betrug laut AMS die Arbeitslosenquote im Jahr 2013 7,6%

Häufigkeitsverteilung bei einem stetigen Merkmal

Stetiges Merkmal X mit vielen unterschiedlichen Ausprägungen

Grundgesamtheit von n Merkmalsträgern

$$x_1, x_2, \ldots, x_n$$

Urliste der Körpergröße von 100 Studenten:

170	183	154	186	165	180	177	178	184	176
178	175	188	153	183	180	165	180	180	196
174	167	181	168	166	187	187	195	164	177
197	161	187	170	171	184	170	187	167	160
162	182	173	168	182	185	155	176	172	164
166	163	170	164	171	173	179	174	180	172
177	175	157	167	165	160	170	173	166	173
166	178	172	176	169	160	169	171	179	179
175	175	190	182	177	185	173	174	172	186
173	173	176	162	174	182	188	188	174	177

Erster Schritt: Ordnen der Daten

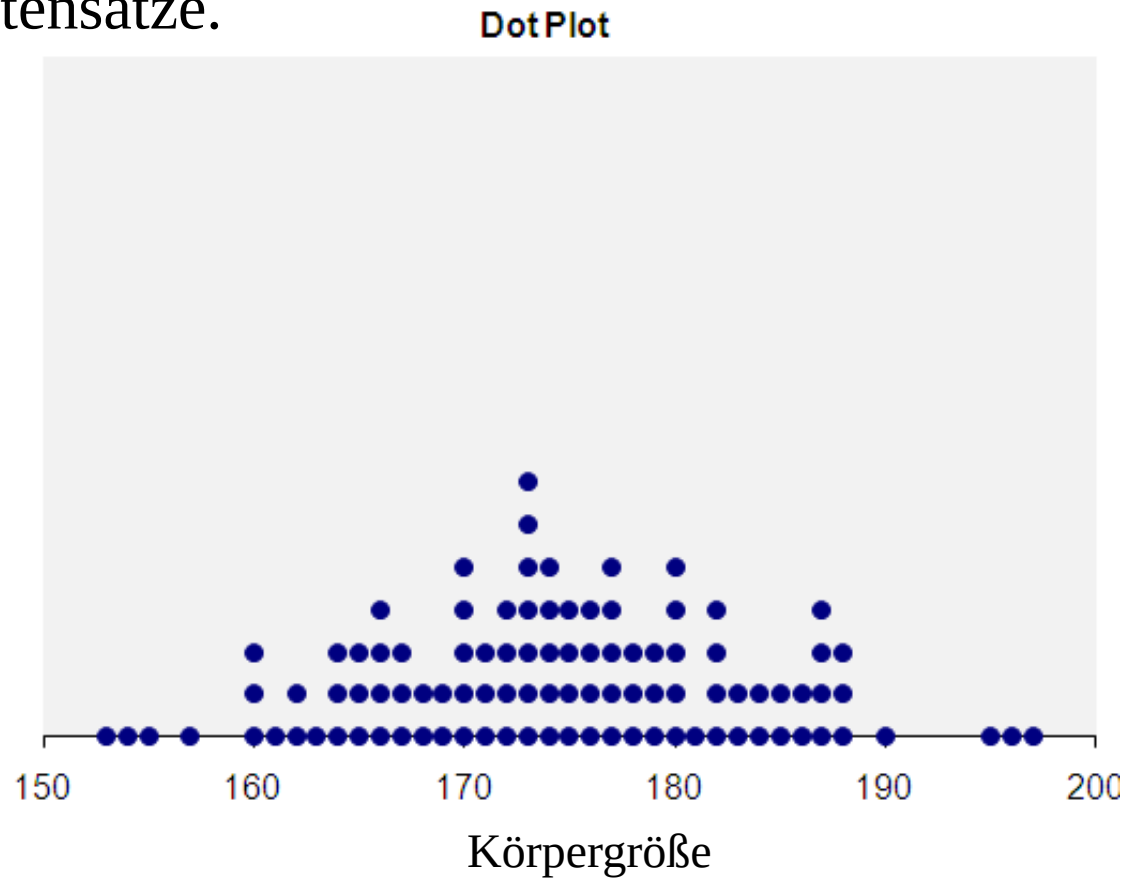
Geordnete Stichpobe: $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$

153	154	155	157	160	160	160	161	162	162
163	164	164	164	165	165	165	166	166	166
166	167	167	167	168	168	169	169	170	170
170	170	170	171	171	171	172	172	172	172
173	173	173	173	173	173	173	174	174	174
174	174	175	175	175	175	176	176	176	176
177	177	177	177	177	178	178	178	179	179
179	180	180	180	180	180	181	182	182	182
182	183	183	184	184	185	185	186	186	187
187	187	187	188	188	188	190	195	196	197

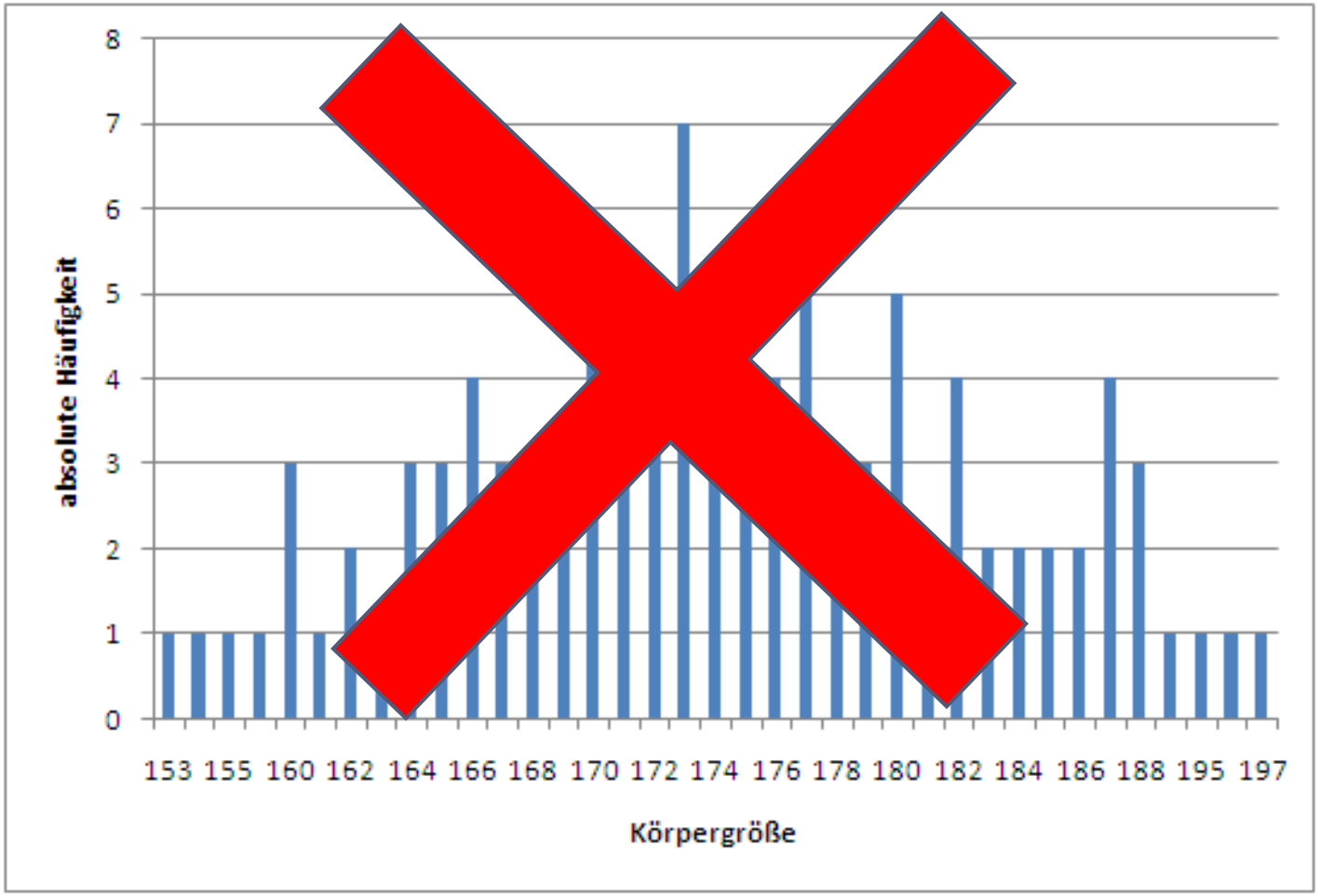
Visualisierung der Verteilung der Größen

Ein Dot-Plot ist eine einfache statistische Graphik für die Darstellung der Verteilung kleinerer bis mittlerer Datensätze.

Visualisierung von Werte-Cluster bzw. von Lücken in der Verteilung sowie für das Aufzeigen extremer Einzelbeobachtungen (Ausreißer).

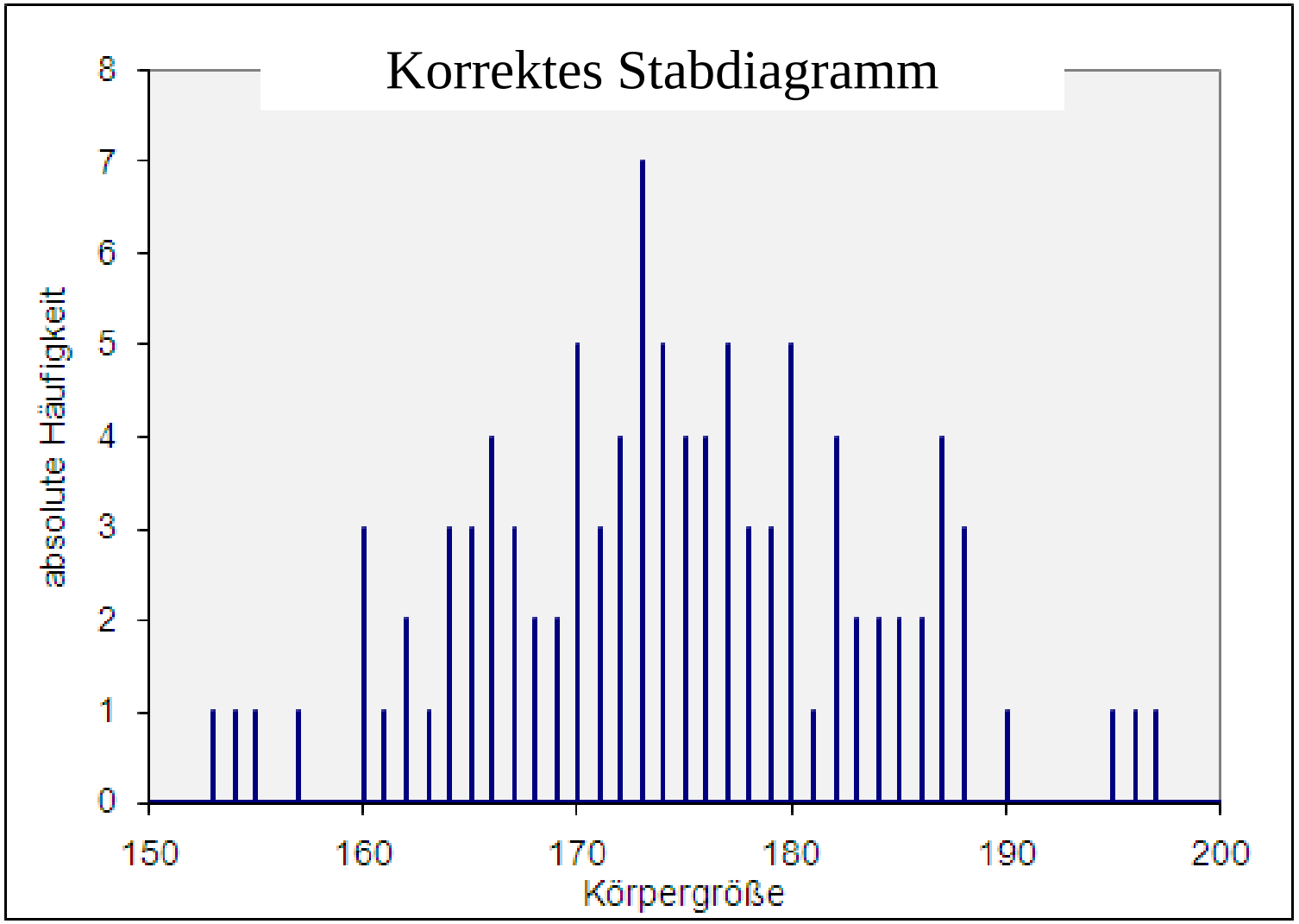


Standardgraphik mit Excel Säulendiagramm



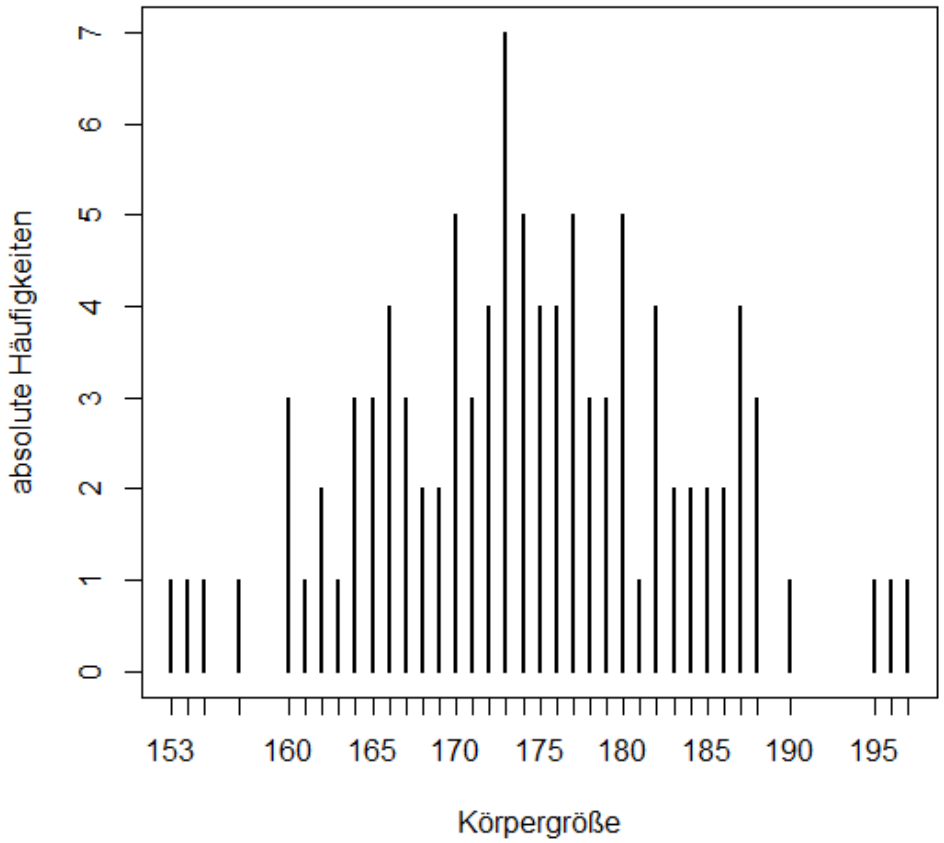
Auf der horizontalen Achse wird die Merkmalsskalierung nicht korrekt widergegeben

Visualisierung der Verteilung der Größen



Stabdiagramm mit R

```
plot(table(Groesse), type="h", xlab= "Körpergröße",
     ylab="absolute Häufigkeiten")
```



Stem & Leaf-Diagram (Stengel-Blatt Diagramm)

N = 100 Median = 174

Spaltenwerte in Einheiten von 10

```
15 : 34
15 : 57
16 : 0001223444
16 : 55566667778899
17 : 00000111222233333344444
17 : 55556666777778888999
18 : 00000122223344
18 : 55667777888
19 : 0
19 : 567
```

Semigraphische Technik:

Systematisches Aufschreiben der Werte, so dass sich ein Bild der Verteilung ergibt

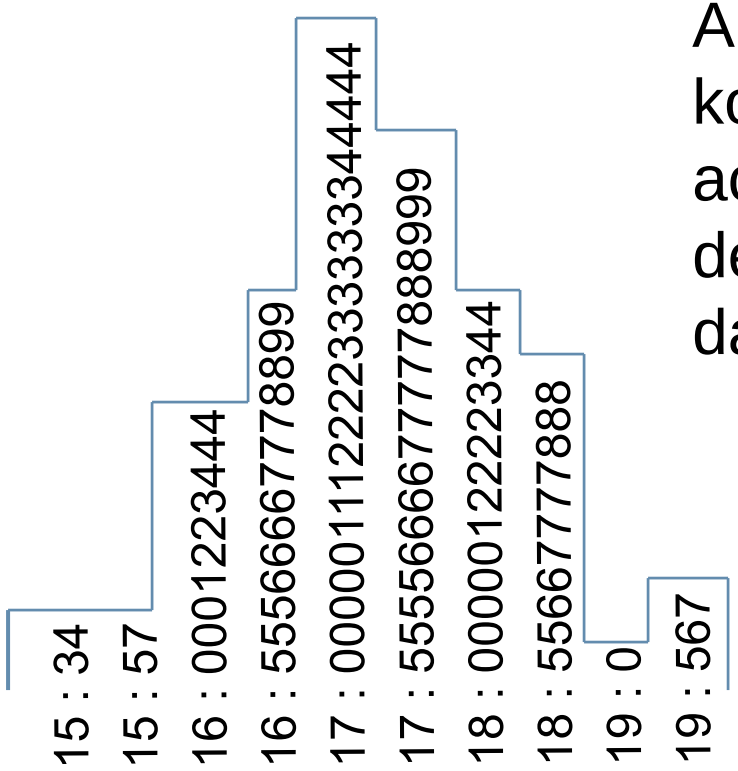
Alle Einzelwerte bleiben exakt erhalten

Nur für kleinere Datensätze geeignet

Stem & Leaf Diagramm mit R

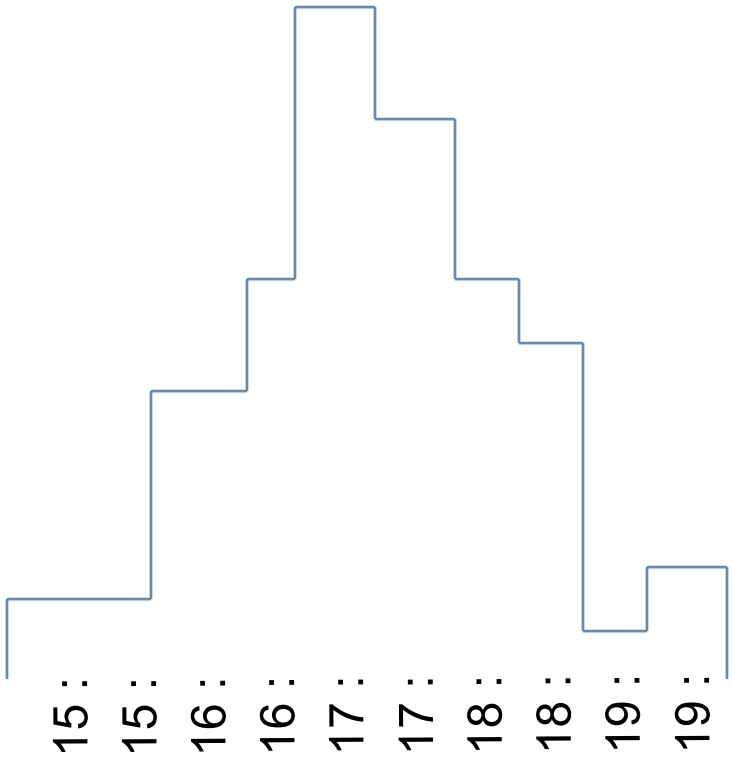
```
> # Analyse des Merkmals Körpergröße
> studenten.df <- read.csv(file="studenten.csv")
> attach(studenten.df)
> stem(Groesse)

  The decimal point is 1 digit(s) to the right of the |

  15 | 34
  15 | 57
  16 | 0001223444
  16 | 55566667778899
  17 | 00000111222233333344444
  17 | 55556666777777888999
  18 | 00000122223344
  18 | 55667777888
  19 | 0
  19 | 567
```

Stemleaf-Diagramm



Abstrahiert man von den konkreten Werten und achtet nur auf die Form der Verteilung, ergibt sich das Histogramm

Histogramm



Das Bild erinnert an ein Säulendiagramm.

Im Gegensatz dazu kommt aber der horizontalen Achse eine geänderte Bedeutung zu.

Numerische Skala!

Häufigkeitstabelle bei stetigen Merkmalen

- Klassierung (Klassifizierung):

  Einteilung des Wertevorrates (Realisationsmöglichkeiten der Merkmalsausprägungen) in nicht überschneidende angrenzende Klassen

  Nach Möglichkeit gleiche Breite !

- Absolute Häufigkeit der Klasse ist die Anzahl der Realisationen mit Werten, die zu dieser Klasse gehören:

  Klasse i sei definiert durch $(u_i,o_i]$

  $$n(u_i<X \leq o_i)=n_i$$

  eindeutige Zuordnung durch halboffene Intervalle

- Relative Häufigkeit der Klasse

  $$h(u_i<X \leq o_i)=n_i/n=h_i$$

Häufigkeitstabelle für klassifizierte Daten

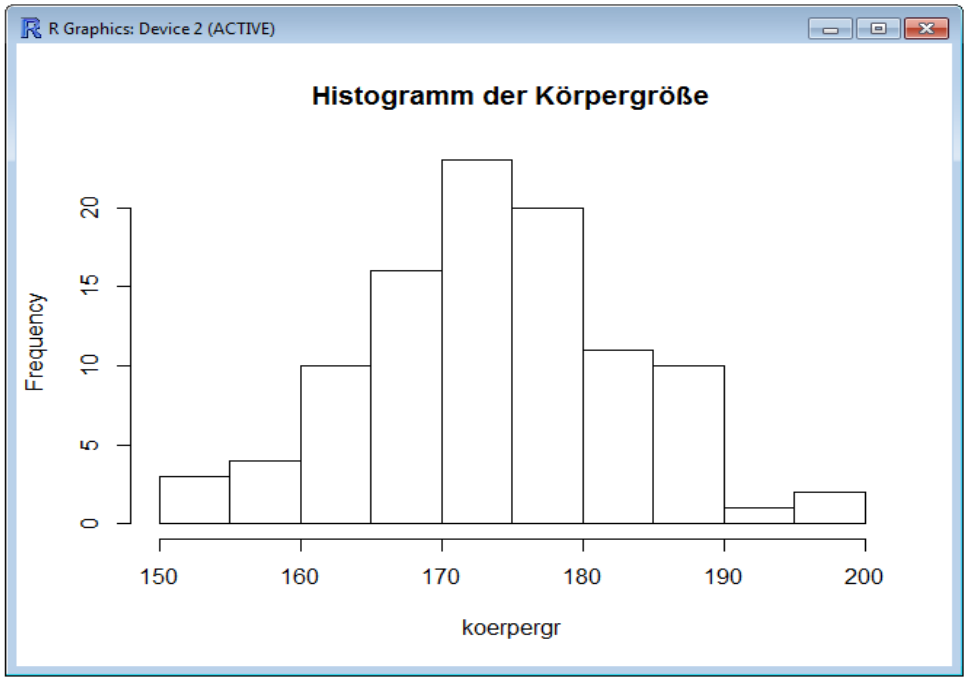
Bereich	n_i	h_i
150+ bis 155	3	0,03
155+ bis 160	4	0,04
160+ bis 165	10	0,10
165+ bis 170	16	0,16
170+ bis 175	23	0,23
175+ bis 180	20	0,20
180+ bis 185	11	0,11
185+ bis 190	10	0,10
190+ bis 195	1	0,01
195+ bis 200	2	0,02
Gesamt	100	1

Histogramm

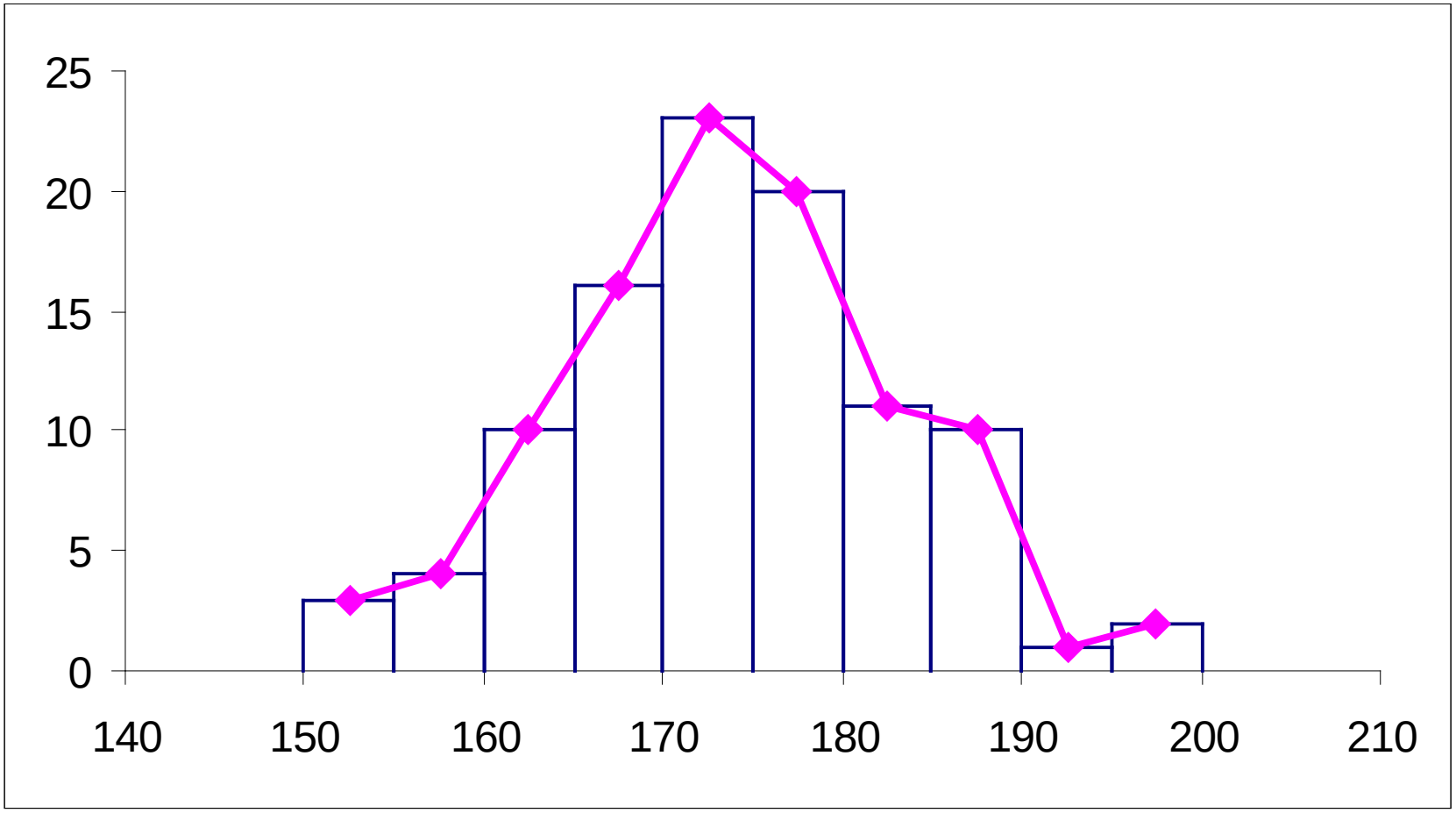
- Ein Histogramm ist die graphische Darstellung einer Häufigkeitstabelle, die sich durch die Klassierung eines stetigen Merkmals in Intervalle ergibt

R Console

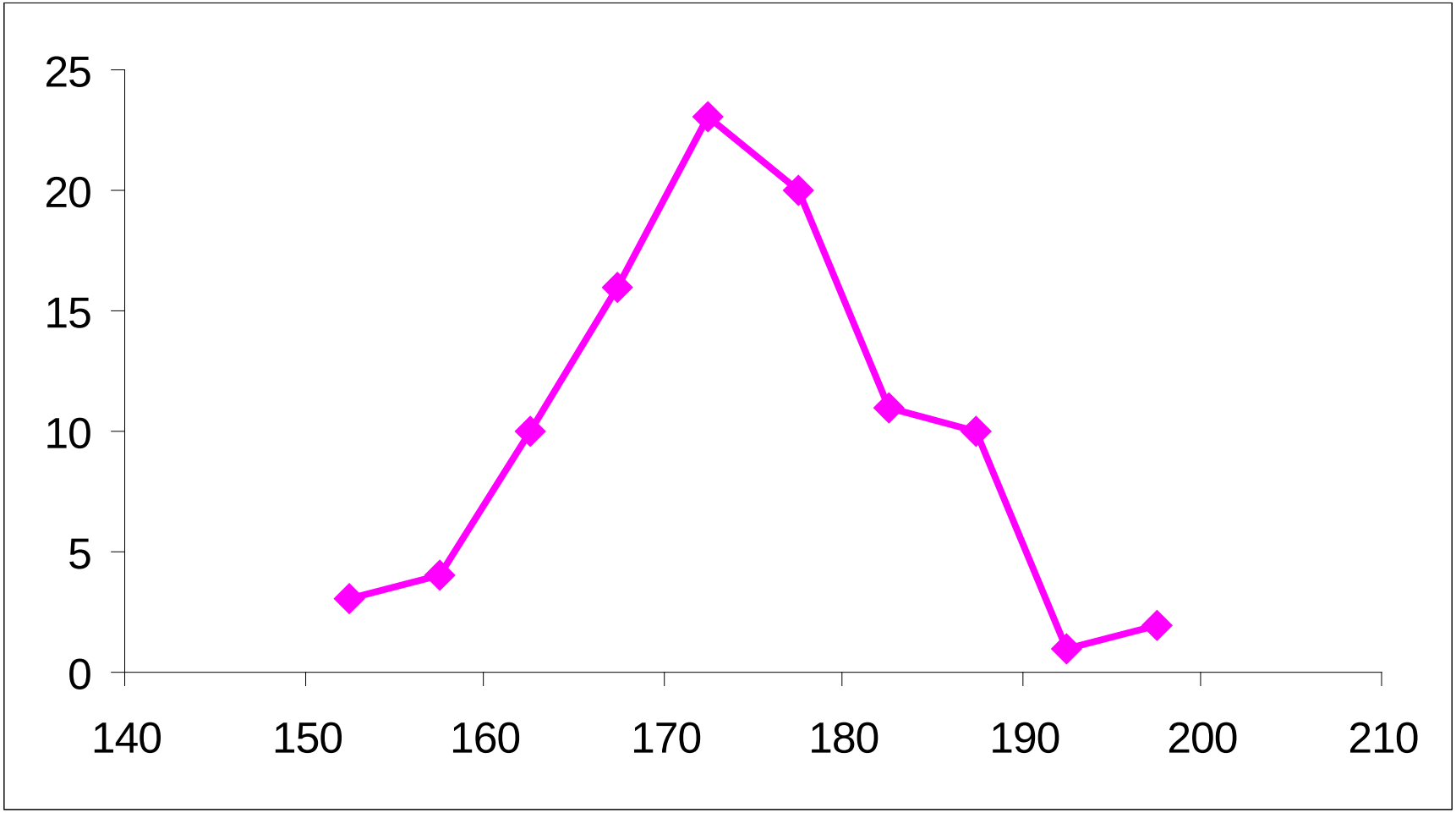
```
> hist(koerpergr, breaks=seq(from=150, to=200, by=5),
+      main="Histogramm der Körpergröße")
>
```



Histogramm mit Polygonzug



Polygonzug



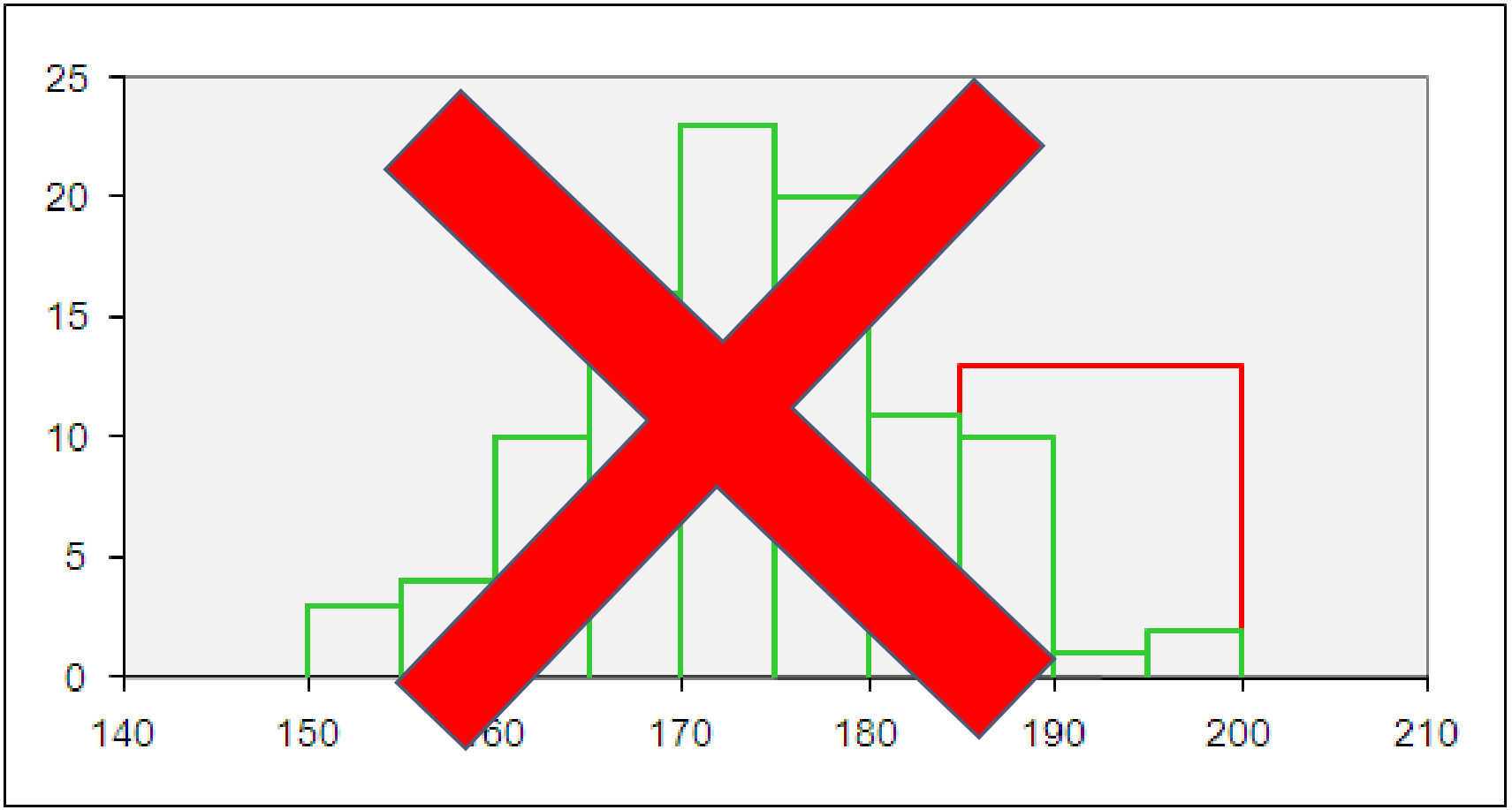
Tabellen in der Praxis

- In der Praxis sind Häufigkeitstabellen oft der Ausgangspunkt einer empirischen Untersuchung (Sekundärdaten)
- Achtung: Klassierte Daten haben nicht 1:1 den selben Informationsgehalt wie die Originaldaten, da die Verteilung innerhalb der Klassen unbekannt ist (vgl. Informationsgehalt von Stem & Leaf-Diagramm und Histogramm)
- Problem offener Klassen:
  z.B.: monatl. Einkommen größer als 100.000,-
- Generell: Ungleich große Klassenbreiten erfordern eine adäquate graphische Darstellung

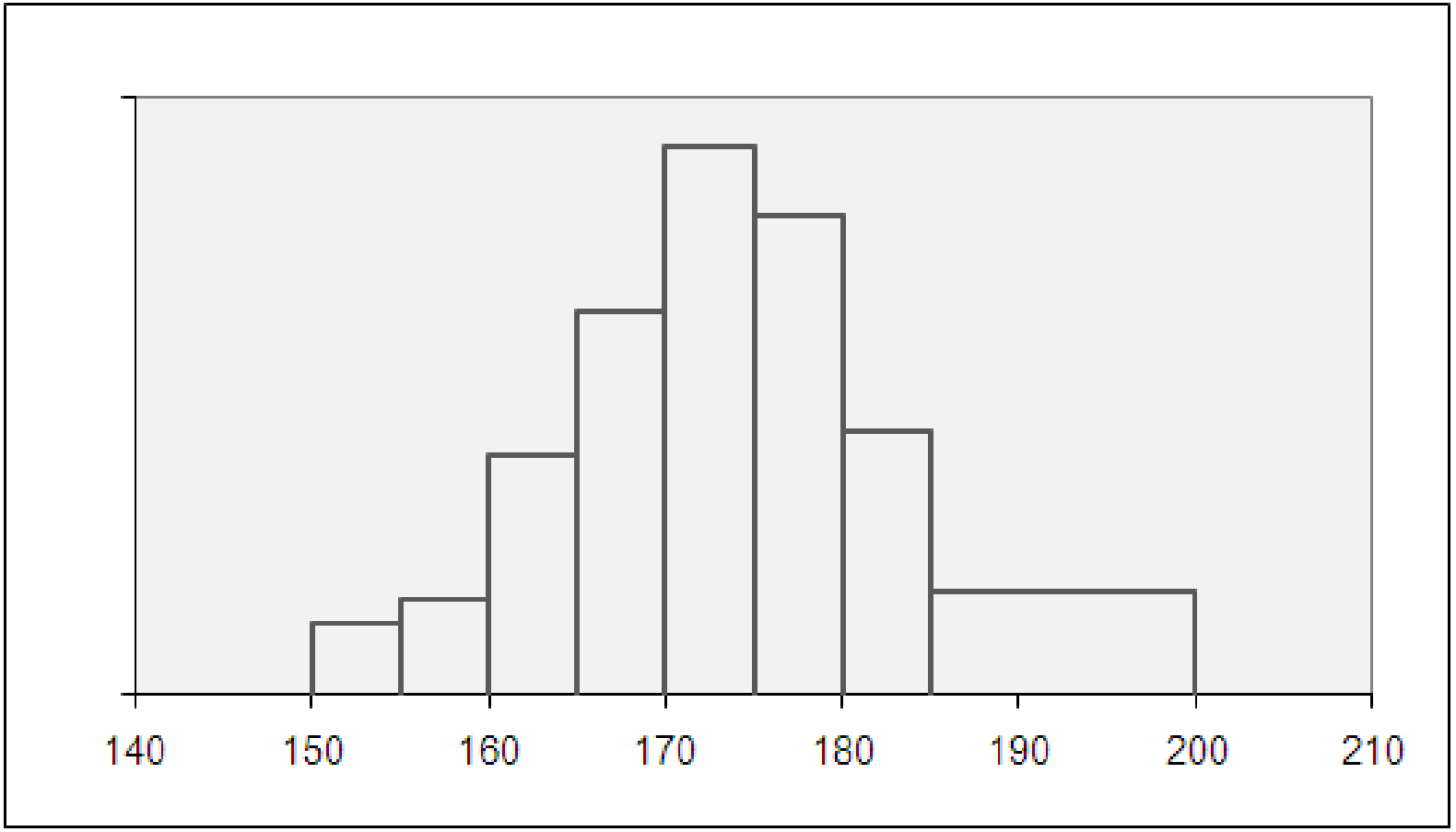
Zusammenfassung von Klassen

Bereich	n_i	h_i
150+ bis 155	3	0,03
155+ bis 160	4	0,04
160+ bis 165	10	0,10
165+ bis 170	16	0,16
170+ bis 175	23	0,24
175+ bis 180	20	0,21
180+ bis 185	11	0,11
185+ bis 200	13	0,13

Fehlerhafte Modifikation des Histogramms

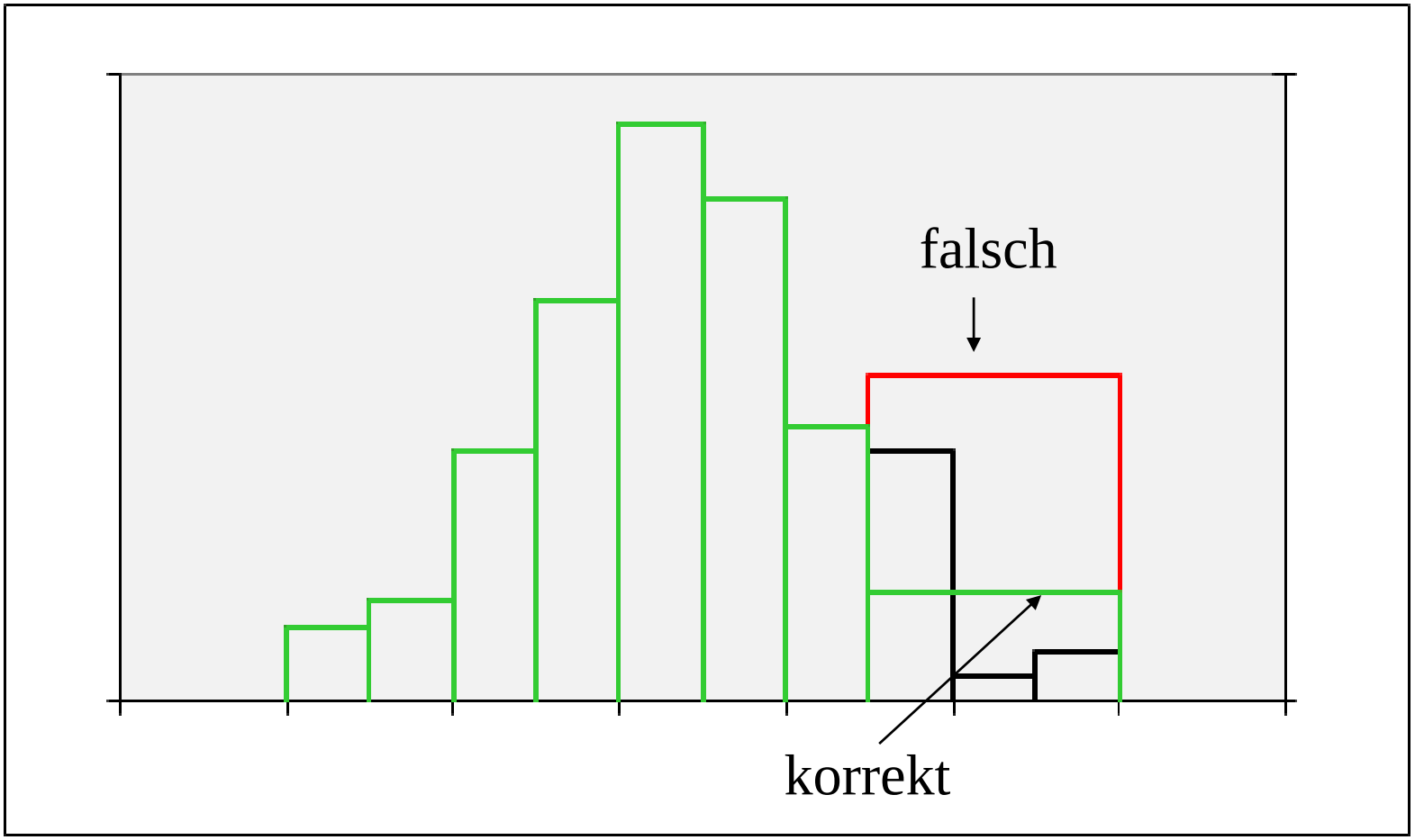


Korrekte Modifikation des Histogramms



Fläche muss konstant bleiben

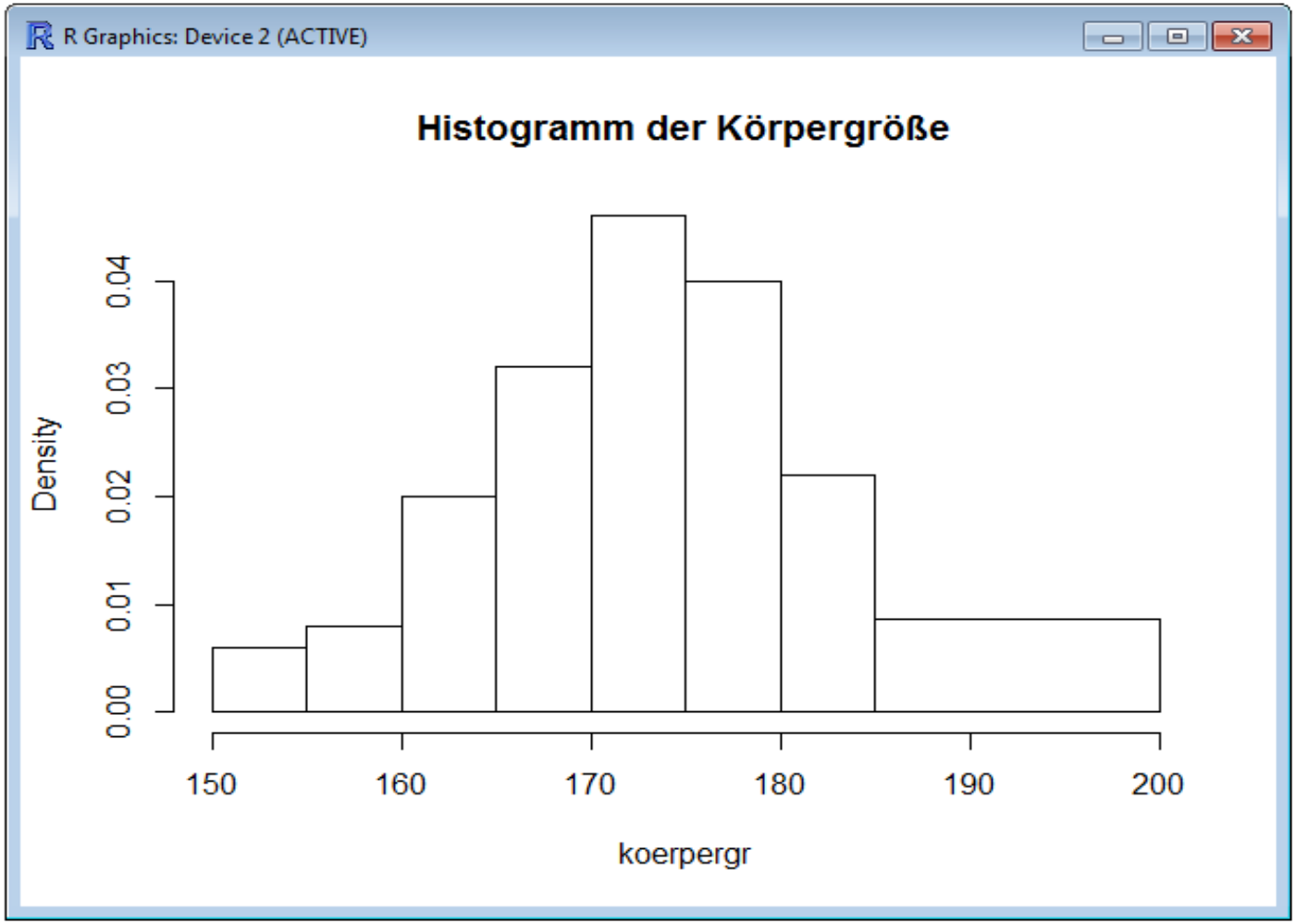
Korrektes Histogramm nach Aggregation



Fläche muss konstant bleiben ➔ Prinzip der Flächentreue

R Console

```
> hist(koerpergr, breaks=c(seq(from=150, to=185, by=5),200),
+      main="Histogramm der Körpergröße")
> 
```



Histogramm <==> Dichtedarstellung

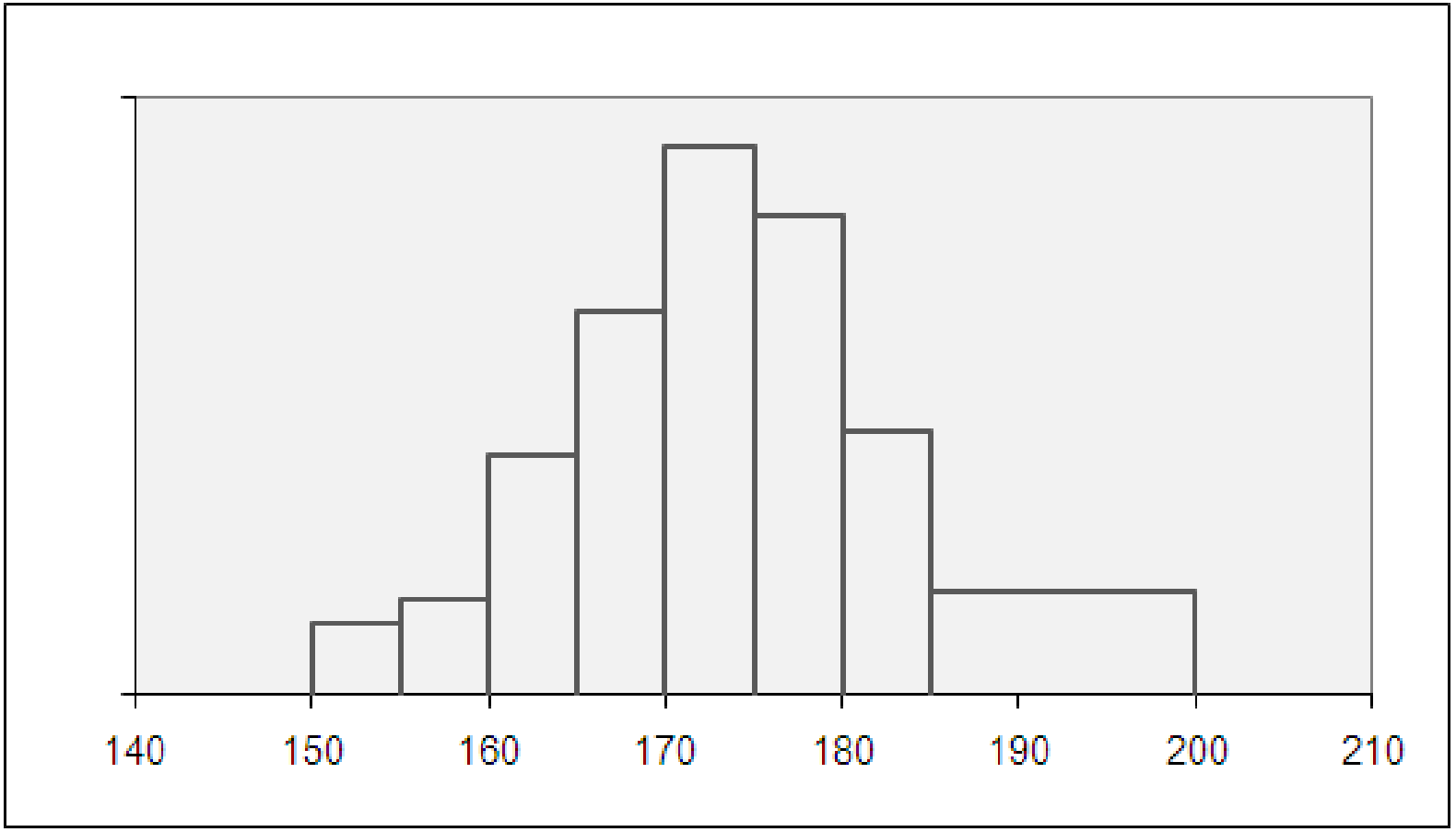
Prinzip der Flächentreue:

- Der Flächeninhalt der Histogramm-Blöcke muss proportional zur Häufigkeit sein
- Die Höhe der Histogramm-Blöcke muss dann die Dichte darstellen
- Dichte ist allgemein eine Häufigkeit bezogen auf eine Einheit
  z.B. Bevölkerungsdichte Anzahl Einwohner je km²
- Häufigkeitsdichte ist die relative Häufigkeit dividiert durch die Klassenbreite
- Fläche = Höhe*Breite und Fläche ~ Häufigkeit
  ➔Höhe muss Häufigkeit/Breite sein
- Bei Klassen konstanter Breite ist diese Unterscheidung für die Visualisierung irrelevant

Berechnung der Häufigkeitsdichte

Bereich	n_i	h_i	b_i	$d_i=h_i/b_i$
150+ bis 155	3	0,030	5	0,006
150+ bis 160	4	0,040	5	0,008
160+ bis 165	10	0,100	5	0,020
165+ bis 170	16	0,160	5	0,032
170+ bis 175	23	0,230	5	0,046
175+ bis 180	20	0,200	5	0,040
180+ bis 185	11	0,110	5	0,022
185+ bis 200	13	0,130	15	0,009
Gesamt	100	1	50	0,183

Korrekte Modifikation des Histogramms



Fläche muss konstant bleiben

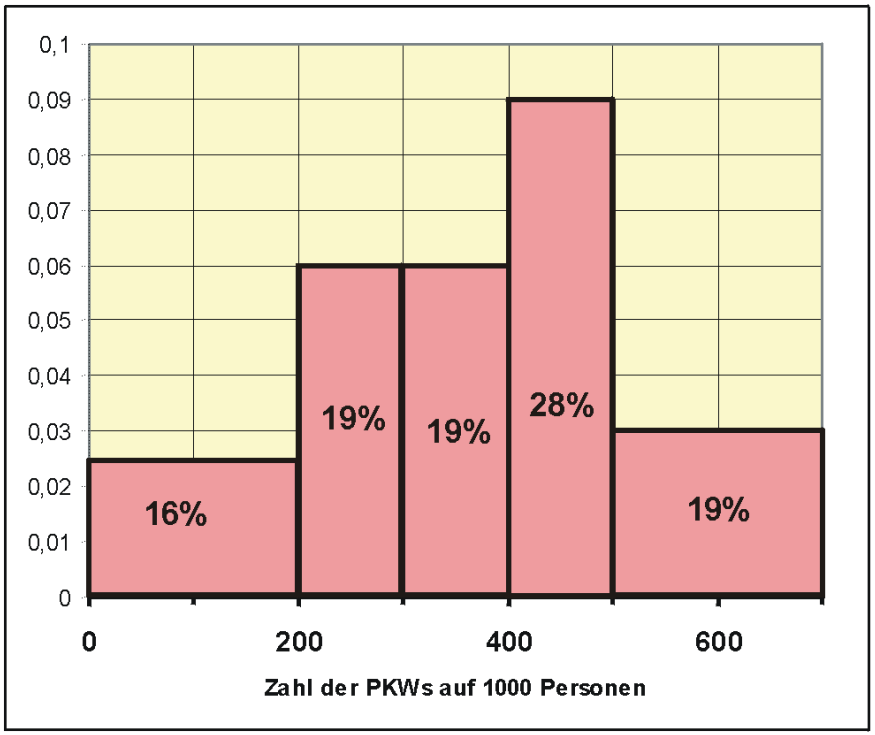
Beispiel: http://de.wikipedia.org/wiki/Histogramm

- Es liegen für 32 europäische Länder als Indikator für den Wohlstand die Zahlen der PKWs pro 1000 Einwohner vor. Die Werte werden wie folgt in Klassen eingeteilt.

Klasse j	Zahl der PKW pro 1000	Anzahl der Länder (absolute Klassenhäufigkeit) n_j	Klassenbreite d_j	Rechteckhöhe (Häufigkeitsdichte) $h_j = n_j/d_j$
1	über 0 - bis 200	5	200 - 0 = 200	0,025
2	über 200 bis 300	6	100	0,06
3	über 300 bis 400	6	100	0,06
4	über 400 bis 500	9	100	0,09
5	über 500 bis 700	6	200	0,03
Summe Σ		32		

Zeichnen von Histogrammen

- Wichtig ist es, dass das Verhältnis der Höhen zueinander korrekt ist.
- Die Bedeutung der Skalierung der y-Achse ist sekundär (siehe dazu das folgende Beispiel)



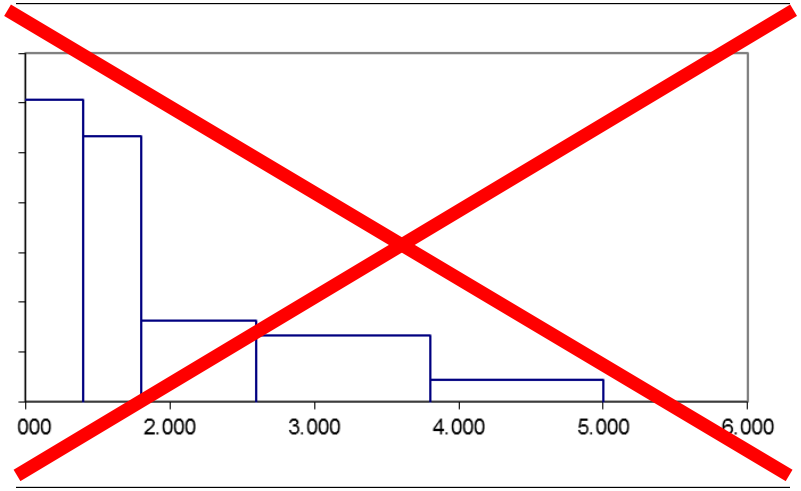
Übungsbeispiel (siehe Excel)

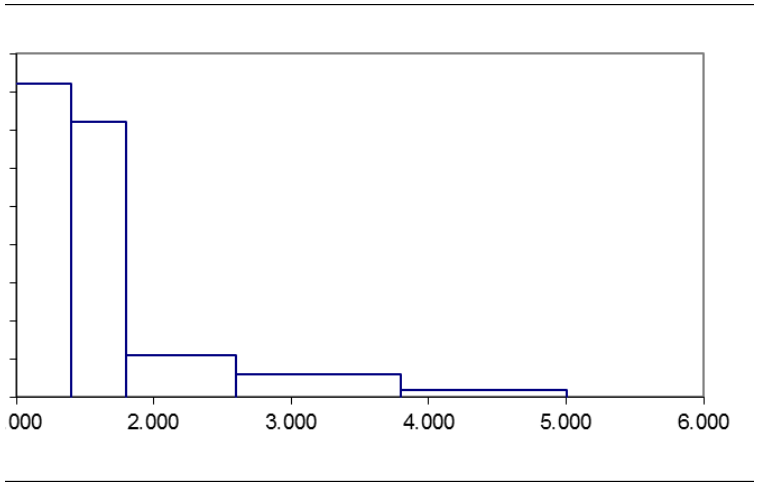
In der nachfolgenden Tabelle sind die Monatsgehälter der Angestellten eines Betriebes angegeben:

Bruttomonatslohn	Anzahl der Angestellten
1.000,-- bis 1.400,-- €	41
1.400,-- bis 1.800,-- €	36
1.800,-- bis 2.600,-- €	11
2.600,-- bis 3.800,--€	9
3.800,-- bis 5.000,--€	3

Monatslohn					
von	bis	Anzahl	relative Häufigkeit	Breite	Dichte per 400€
1.000	1.400	41	0,4100	400	0,4100
1.400	1.800	36	0,3600	400	0,3600
1.800	2.600	11	0,1100	800	0,0550
2.600	3.800	9	0,0900	1.200	0,0300
3.800	5.000	3	0,0300	1.200	0,0100
		100	1,0000		

Skizziere ein korrektes Histogramm für die Häufigkeitsverteilung





Pragmatische Vorgangsweise

- Um den Rechenaufwand zu minimieren haben wir als Einheit der Klassenbreite 400€ gewählt.
- Dann können wir bei den Klassen mit Breite 400€ einfach die relative Häufigkeit direkt als Dichte im Histogramm verwenden.
- Die relative Häufigkeiten der Klassen mit Breite 800 müssen dann durch 2 dividiert werden.
- Die relative Häufigkeiten der Klassen mit Breite 1.200 müssen dann durch 3 dividiert werden.

Wahl der Klassen für ein Histogramm

- In der Praxis ist die Wahl der Klassengrenzen bzw. der Klassenanzahl relativ willkürlich.
- In der Praxis wählt man die Klassenanzahl je nach Datenlage und Fragestellung zwischen 5 – 20

Faustregeln

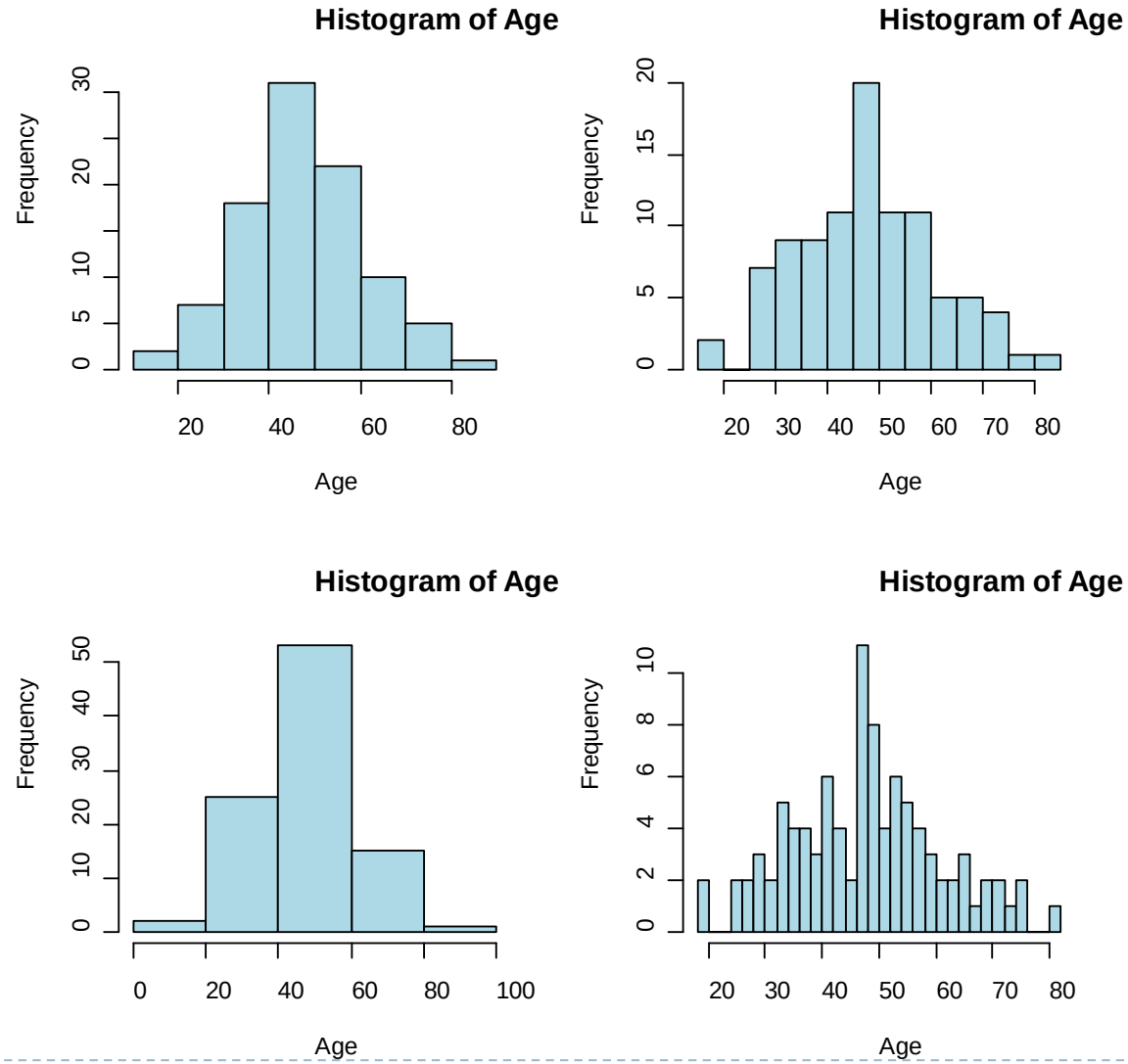
Anzahl der Messungen	Balkenzahl
<50	5 bis 7
50 bis 100	6 bis 10
100 bis 250	7 bis 12
>250	10 bis 20

- Annäherung an die Klassenbreite: Spannweite (Differenz zwischen größtem und kleinsten Wert) bzw. durch gewünschte Klassenanzahl dividieren
- Weitere Aspekte:
  - Klassenmitten sollten mit beobachteten Werten übereinstimmen
  - Wahl „schöner“ Grenzen im Sinne des dekadischen Systems

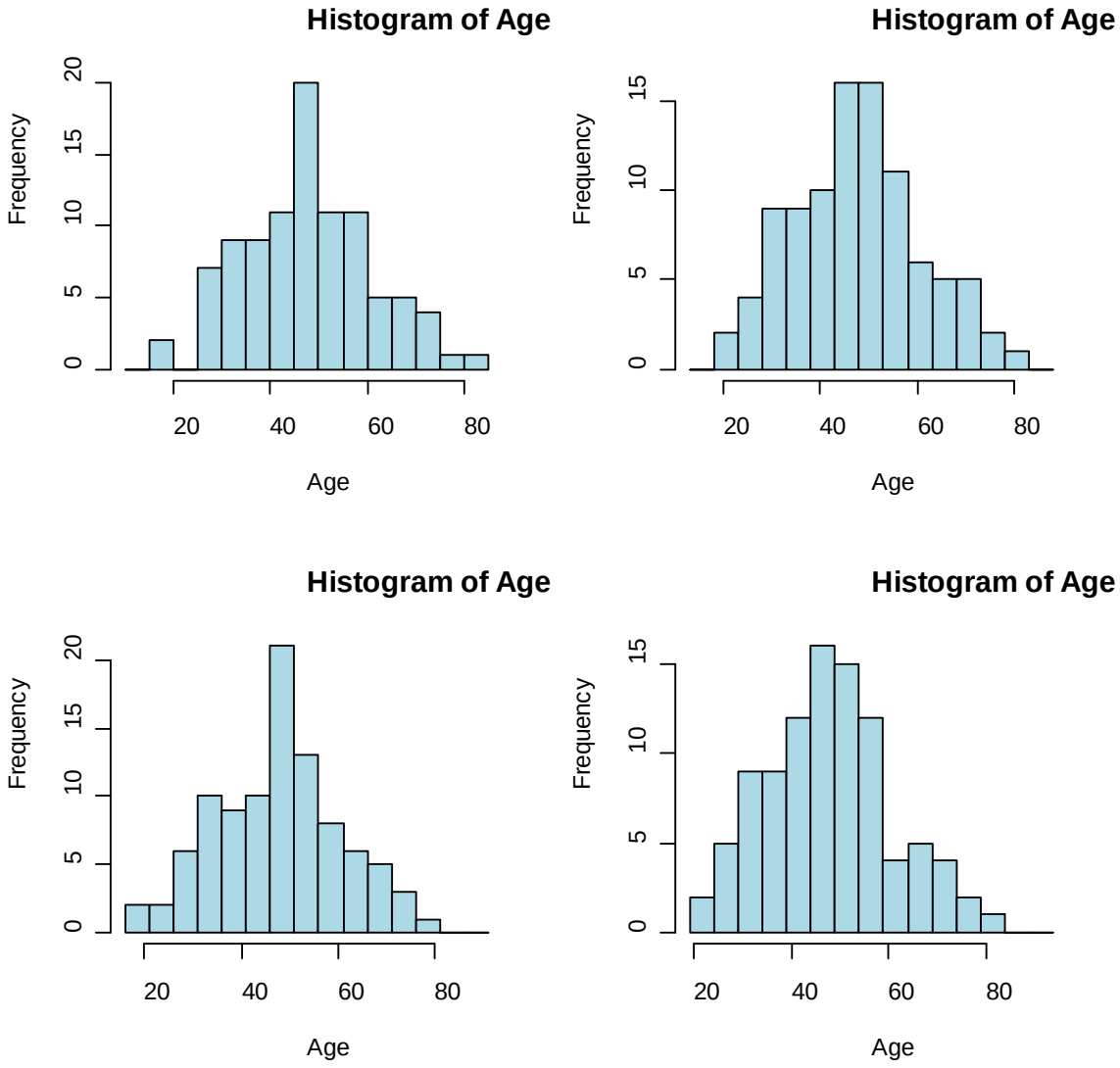
Kritik am klassischen Histogramm

- Das resultierende Bild von der Gestalt der Verteilung hängt von der letztlich willkürlichen Wahl der Anzahl der Intervalle und vom Startpunkt der Intervallbildung ab

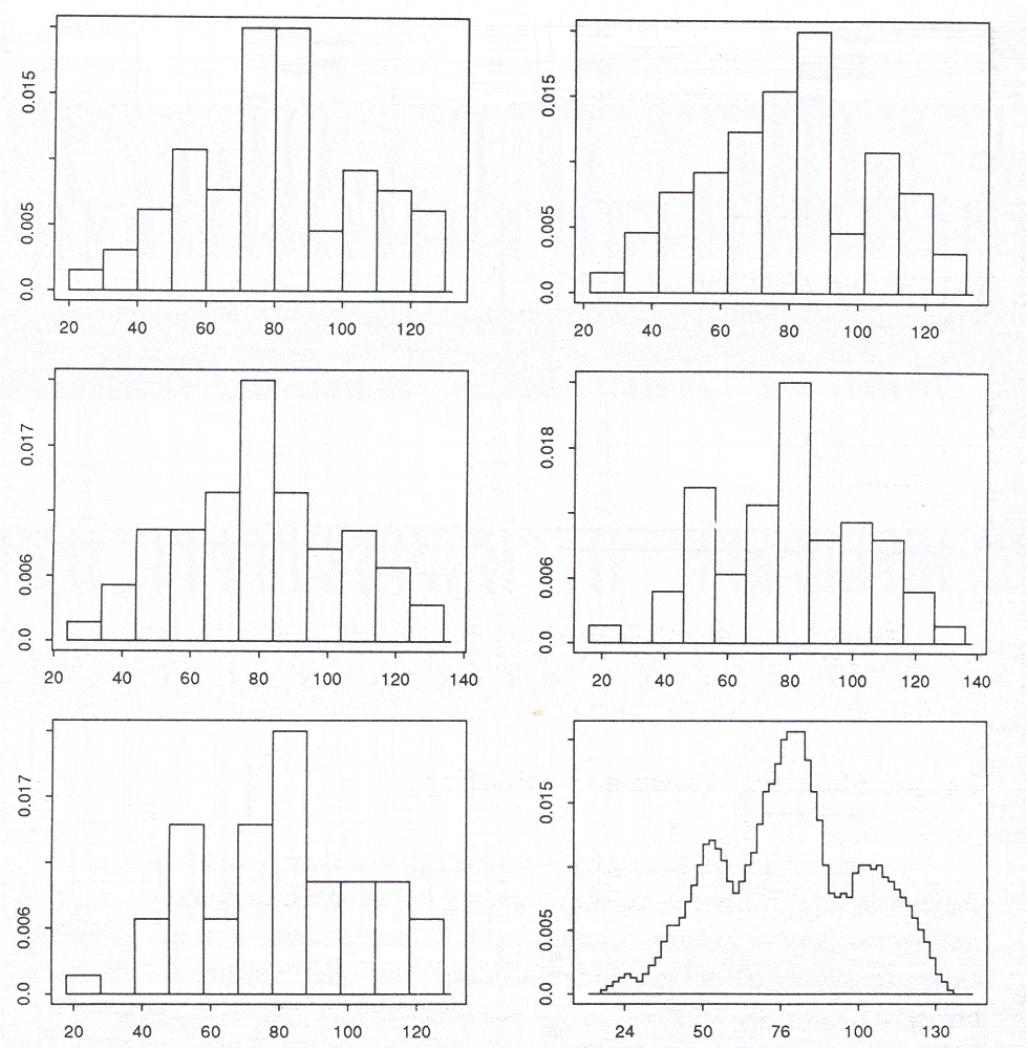
Beispiel: Altersverteilung einer Kohorte



Verschieben der Klassen-Intervalle

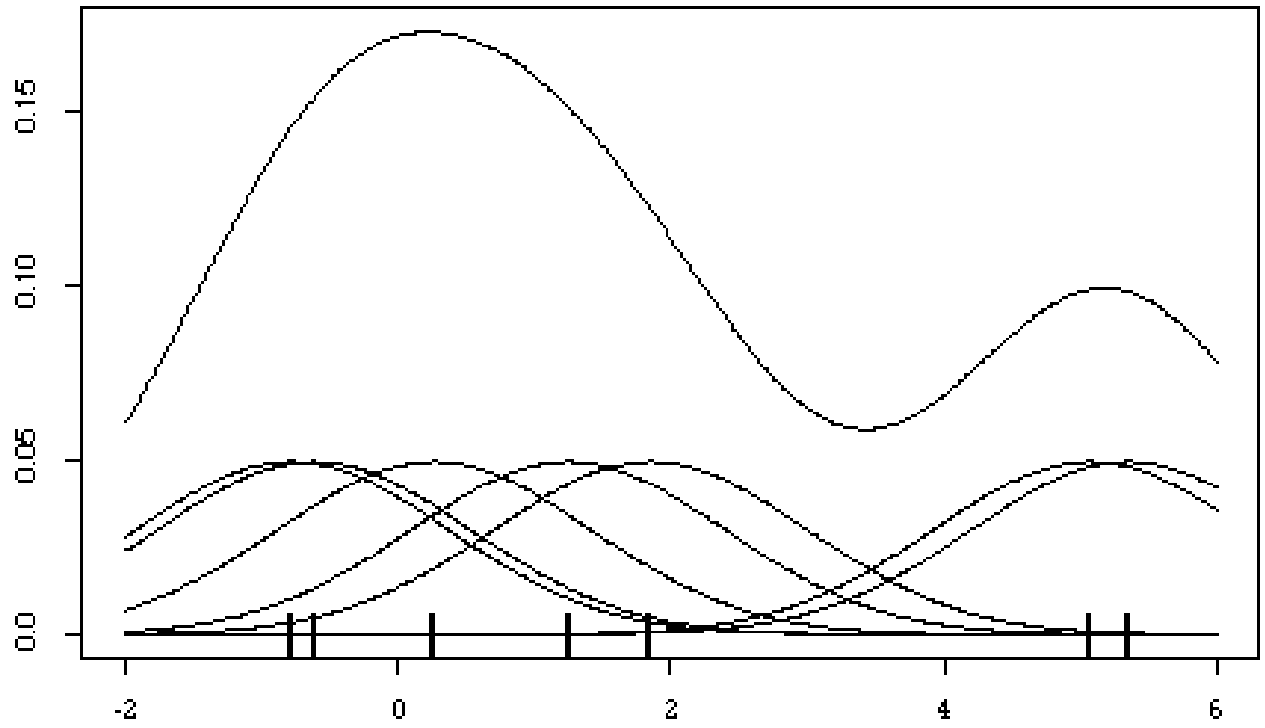


Average Shifted Histogram



ASH (Scott 1985)
Durch Mittelung über mehrere Histogramme, die mit einem unterschiedlichen Startwert für die Klassenbildung erzeugt werden, erhält man ein reliableres Schaubild der Verteilung

Alternative: Kernschätzer

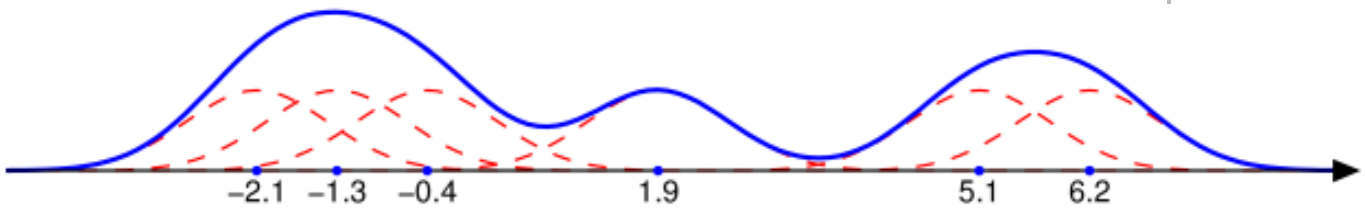


Ein Kern bezeichnet eine symmetrische Funktion, die rund um den empirischen Datenpunkt aufgetragen wird.
Das Bild der Verteilung ergibt sich dann durch die Summe über alle Funktionswerte im gesamten Bereich der x-Achse

Dichte-Schätzung mit Kernfunktionen

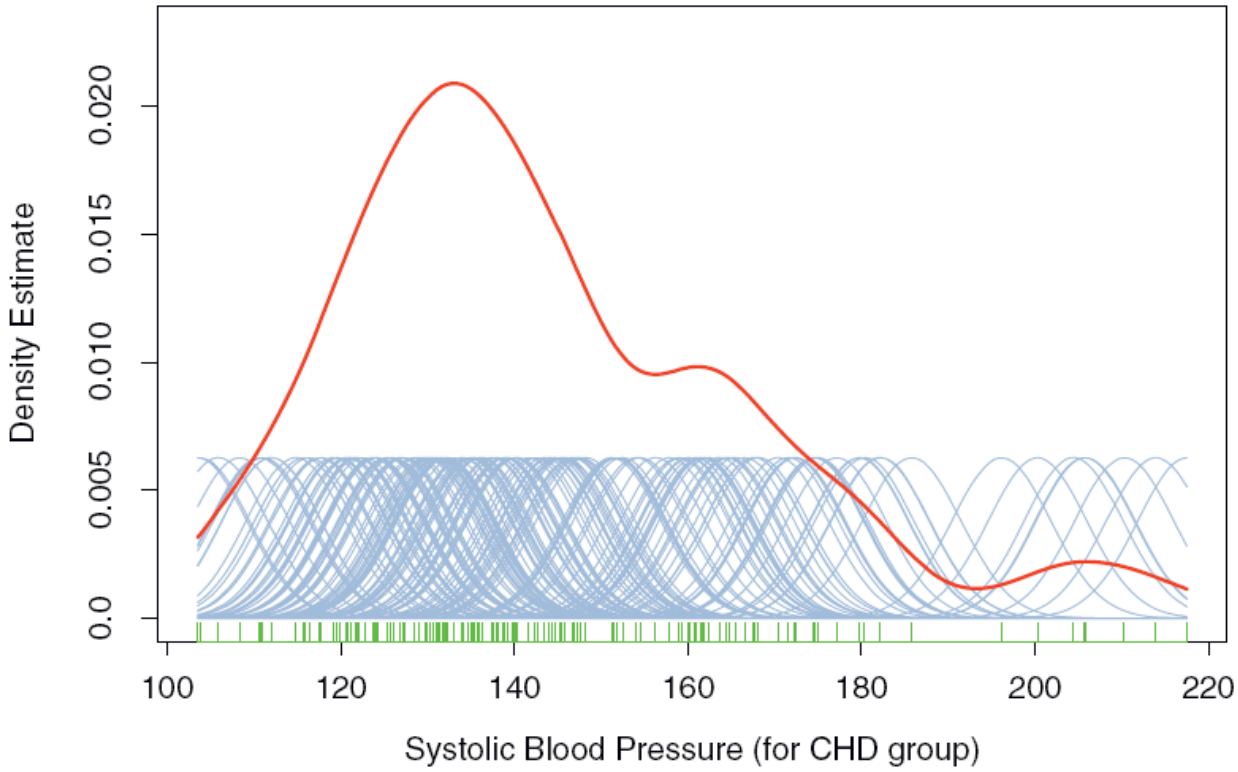
- Kernel density estimation or **Parzen window** method is an alternative way of graphing the shape of a distribution.
- A Kernel is a symmetric function putting mass around any data point x_i
- Various types of Kernels are used (Gaussian, triangular, rectangular)
- The adequate choice of the bandwidth h is critical

$$\widehat{f}_h(x) = \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{x - x_i}{h}\right) \qquad K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}.$$



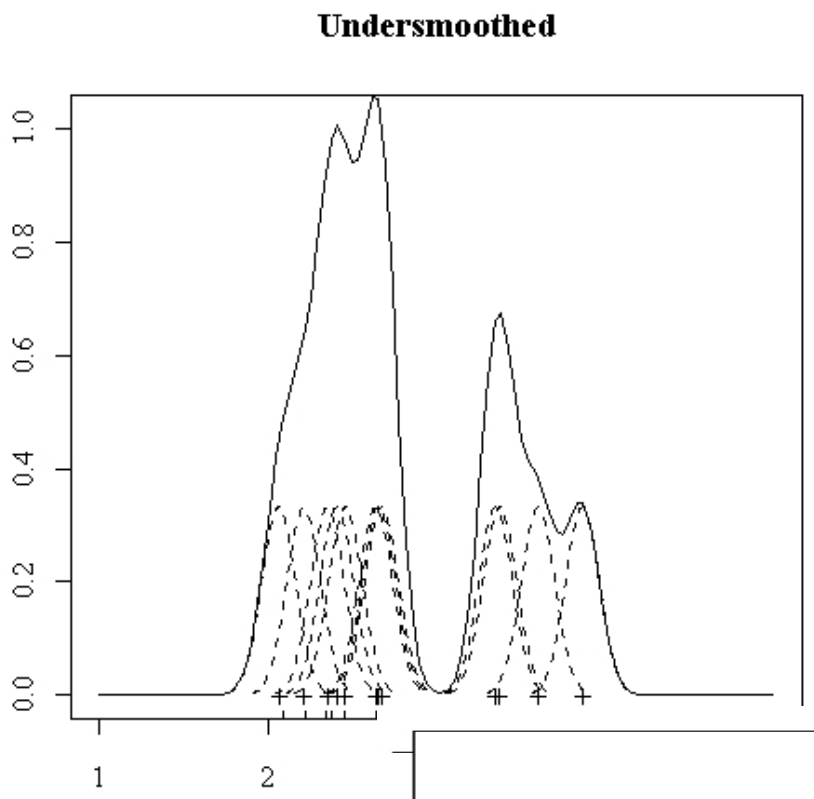
Kern Prinzip für Dichteschätzung

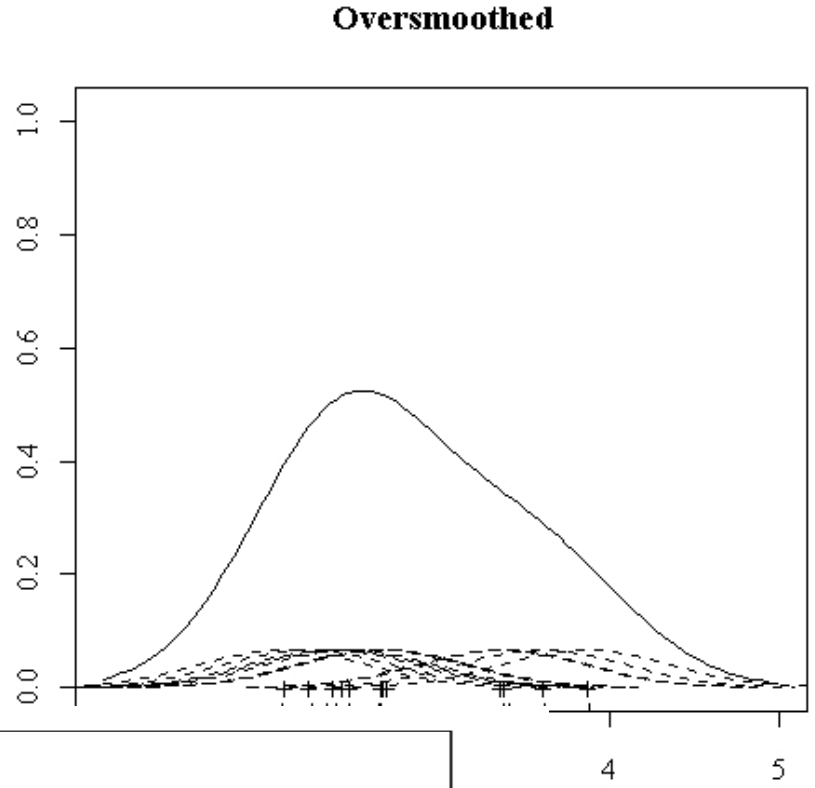
Für jeden Datenpunkt (grün) wird eine Kernfunktion (blaue Linien) geschätzt und aus der Summation dieser Kernfunktionen ergibt sich die geschätzte Kerndichtefunktion (rote Linie).

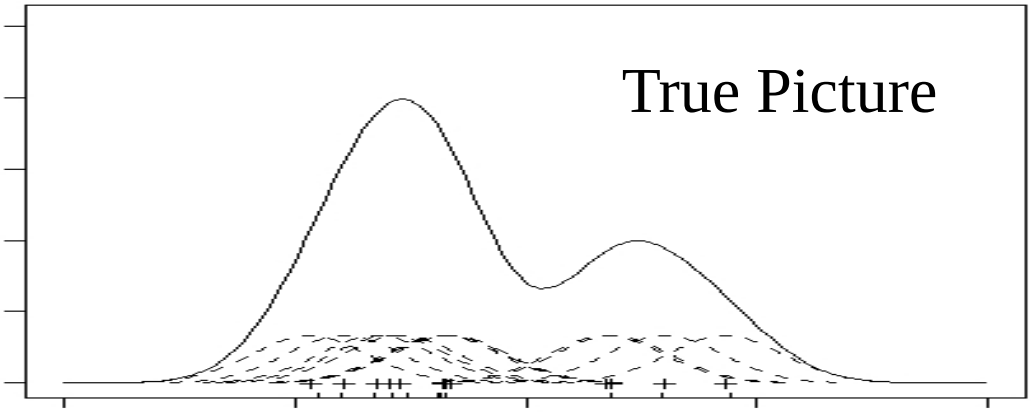


Quelle: Elements of Statistical Learning (2009), Trevor Hastie, Robert Tibshirani, Jerome Friedman, Springer, 2. Auflage, S. 208

Selektion der Bandweite (Simulation)



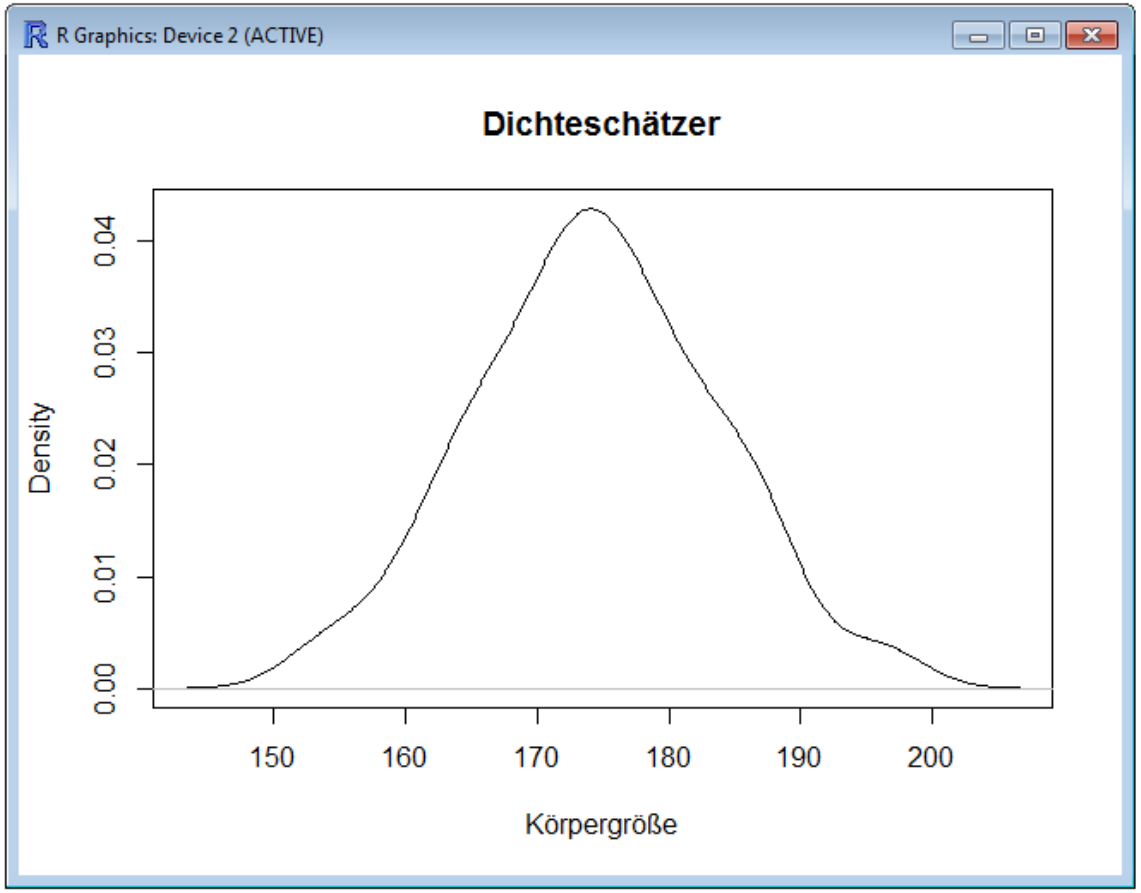




Kernschätzer für Studentendaten mit Gauss-Kern

R Console

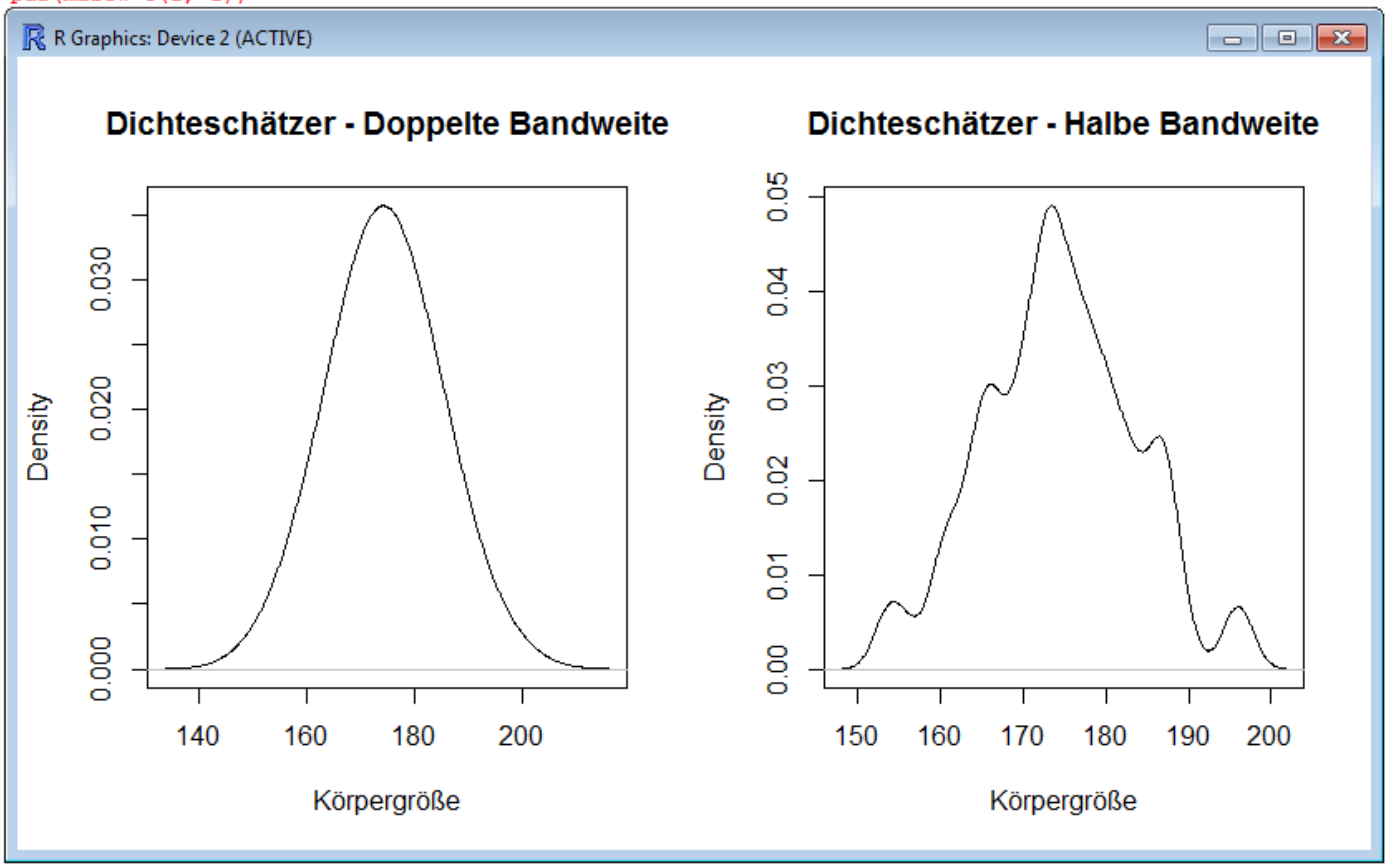
```
> d <- density(koerpergr)
> plot(d, xlab= "Körpergröße", main="Dichteschätzer")
>
```



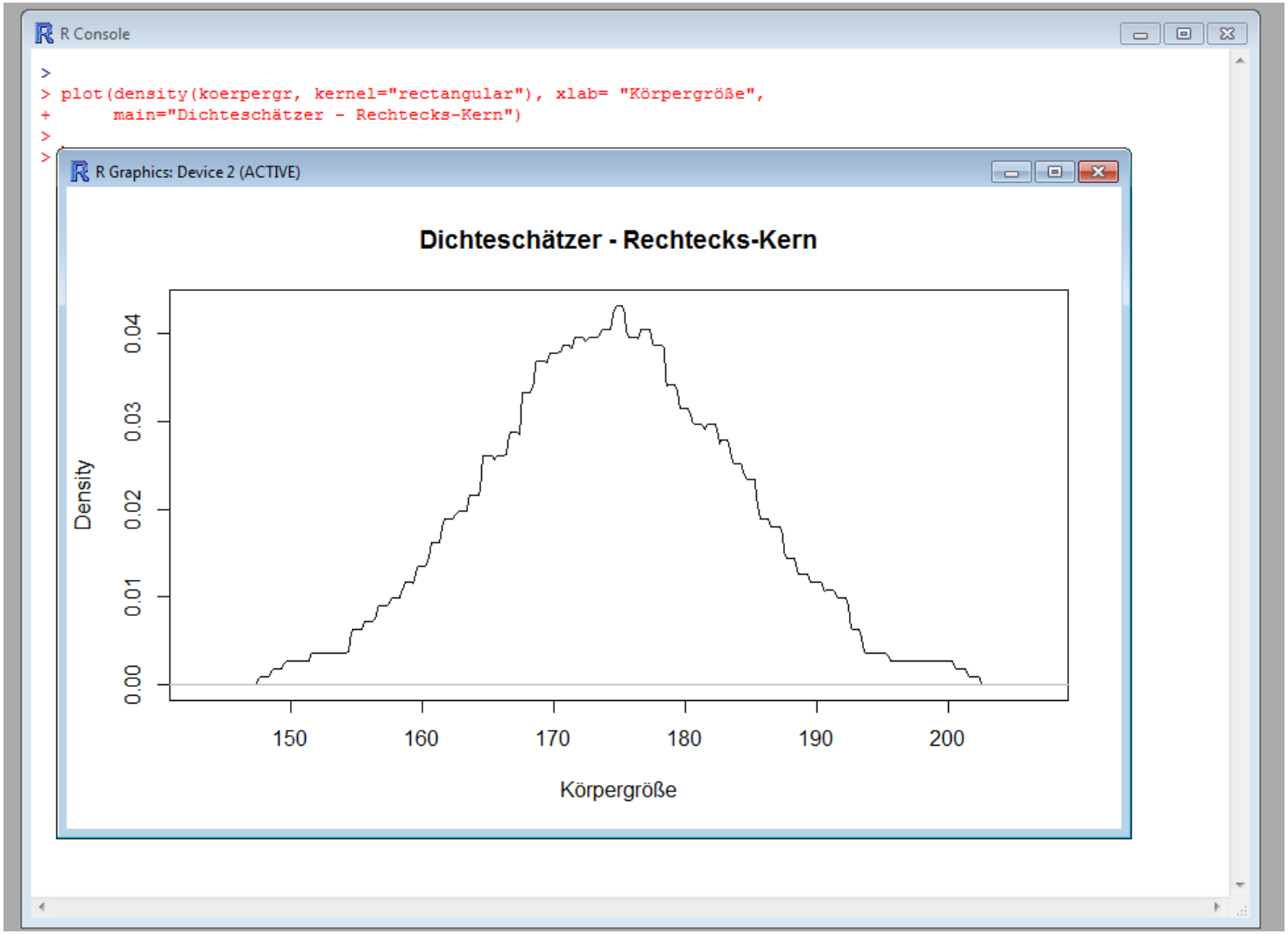
Variation der Bandweite

R Console

```
> par(mfrow=c(1, 2))
> plot(density(koerpergr, adj=2), xlab= "Körpergröße",
+      main="Dichteschätzer - Doppelte Bandweite")
> plot(density(koerpergr, adj=0.5), xlab= "Körpergröße",
+      main="Dichteschätzer - Halbe Bandweite")
> par(mfrow=c(1, 1))
>
```

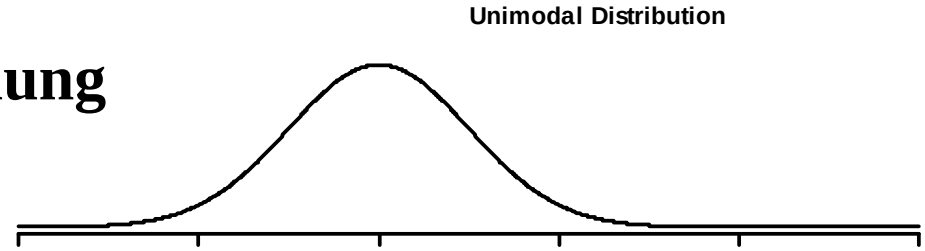


Andere Kernfunktion

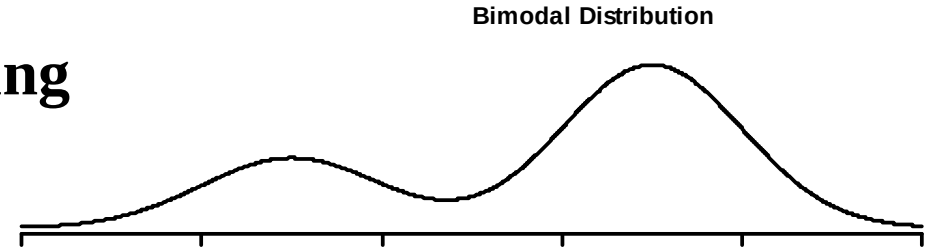


Typisierung von Verteilungen

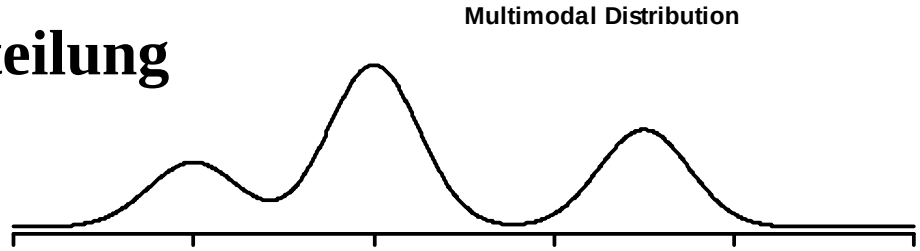
Unimodale Verteilung



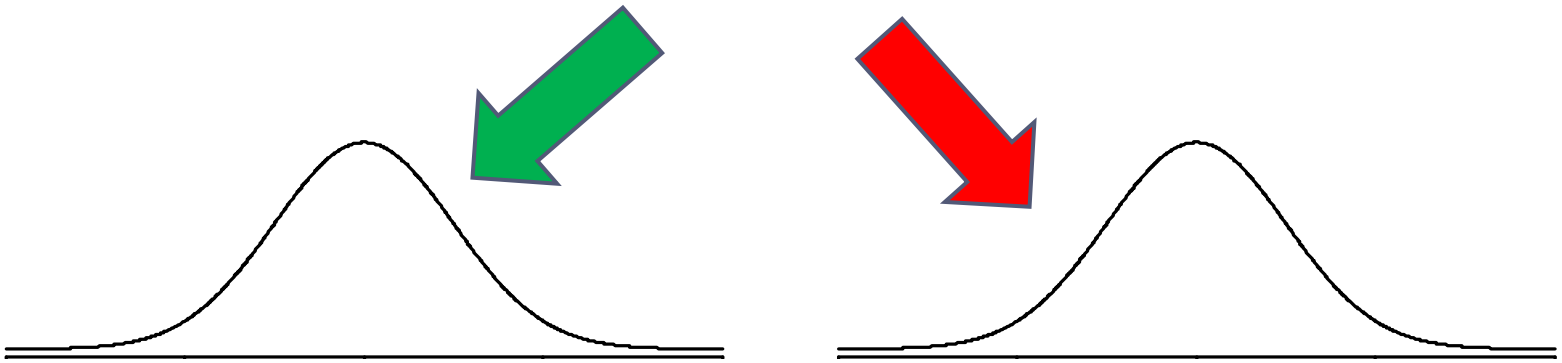
Bimodale Verteilung



Multimodale Verteilung



Schiefe einer Verteilung



Unimodale symmetrische Verteilung

rechtsschief

linksschief

Kumulierte Häufigkeiten

- Hinweis:
  Damit das Kumulieren inhaltlich sinnvoll ist, muss das Merkmal zumindest ordinal skaliert sein!
- Oft ist man nicht an der Häufigkeit einzelner Merkmalsausprägungen interessiert, sondern an der Häufigkeit des Vorkommens von Intervallen.
- Typische Fragestellung:
  - Wie groß ist der Anteil aller Merkmalsträger mit einem Merkmalswert größer (bzw. kleiner) als ein bestimmter Wert x?
- Zur einfachen Beantwortung summiert man die Häufigkeitstabelle schrittweise auf.

Kumulierte Häufigkeiten bei diskreten Merkmalen

Beispiel: "Produktives Denken"

i	x_i	n_i	N_i	h_i	h_i in %	H_i	H_i in %
1	0	0	0	0,00	0,00%	0,00	0,00%
2	1	0	0	0,00	0,00%	0,00	0,00%
3	2	0	0	0,00	0,00%	0,00	0,00%
4	3	7	7	0,06	5,83%	0,06	5,83%
5	4	12	19	0,10	10,00%	0,16	15,83%
6	5	38	57	0,32	31,67%	0,48	47,50%
7	6	29	86	0,24	24,17%	0,72	71,67%
8	7	27	113	0,23	22,50%	0,94	94,17%
9	8	6	119	0,05	5,00%	0,99	99,17%
10	9	1	120	0,01	0,83%	1,00	100,00%
Gesamt		120		1,00	100,00%		

Kumulierte Häufigkeiten

- Die absoluten kumulierten Häufigkeiten geben die Anzahl der Beobachtungen an, die einen bestimmten Wert x nicht übertreffen.

$$N(X \leq x)$$

z.B. 19 Personen haben einen Score-Wert kleiner gleich 4

- Die entsprechenden relativen kumulierten Häufigkeiten bezeichnen wir mit

$$H(X \leq x) = N(X \leq x)/n$$

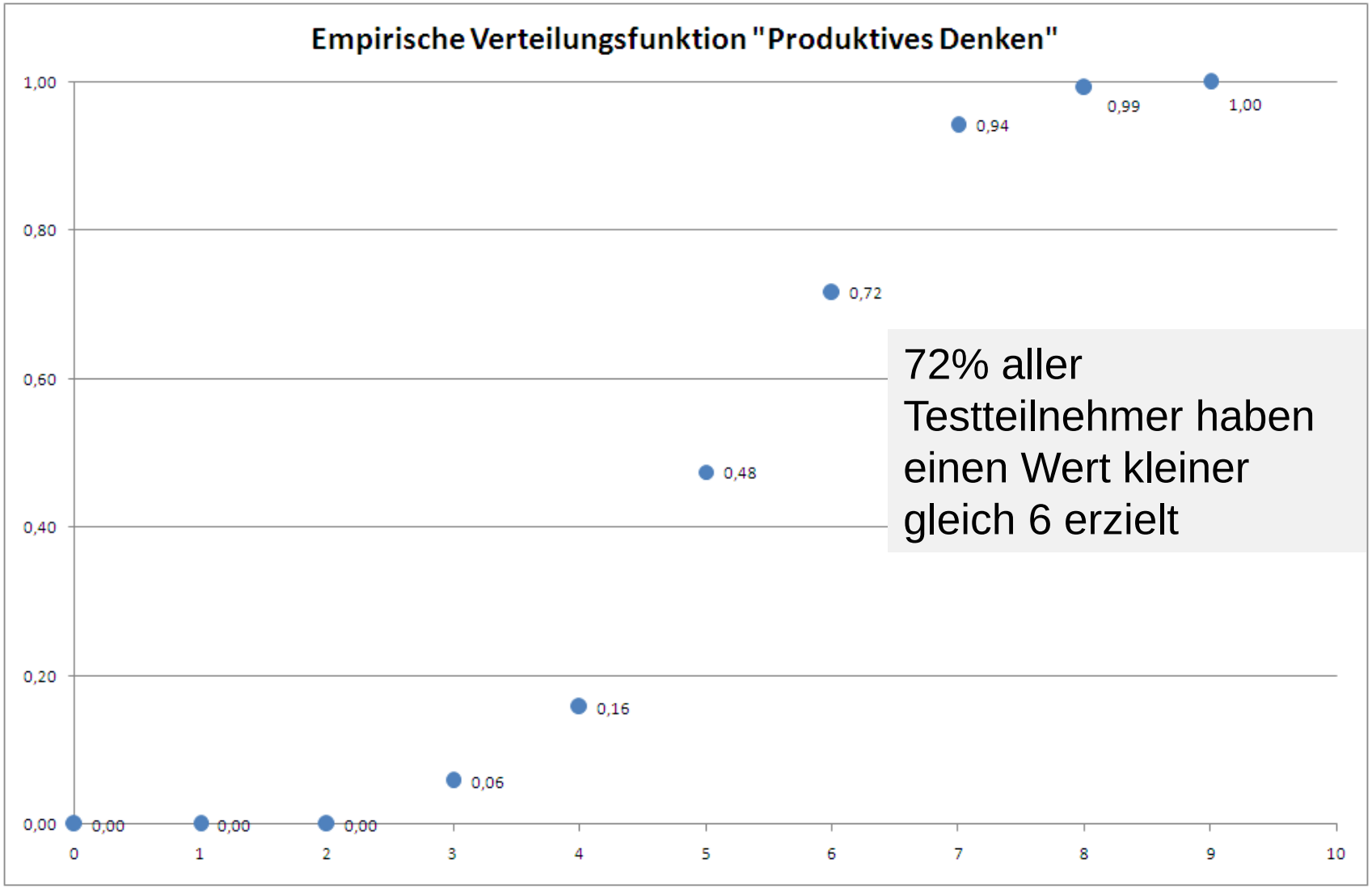
z.B. 15,8% der Personen haben einen Score-Wert kleiner gleich 4

Sie geben uns den Anteil der Beobachtungen mit einem Wert kleiner gleich x an.

- Die empirische Verteilungsfunktion F(x) ist definiert durch

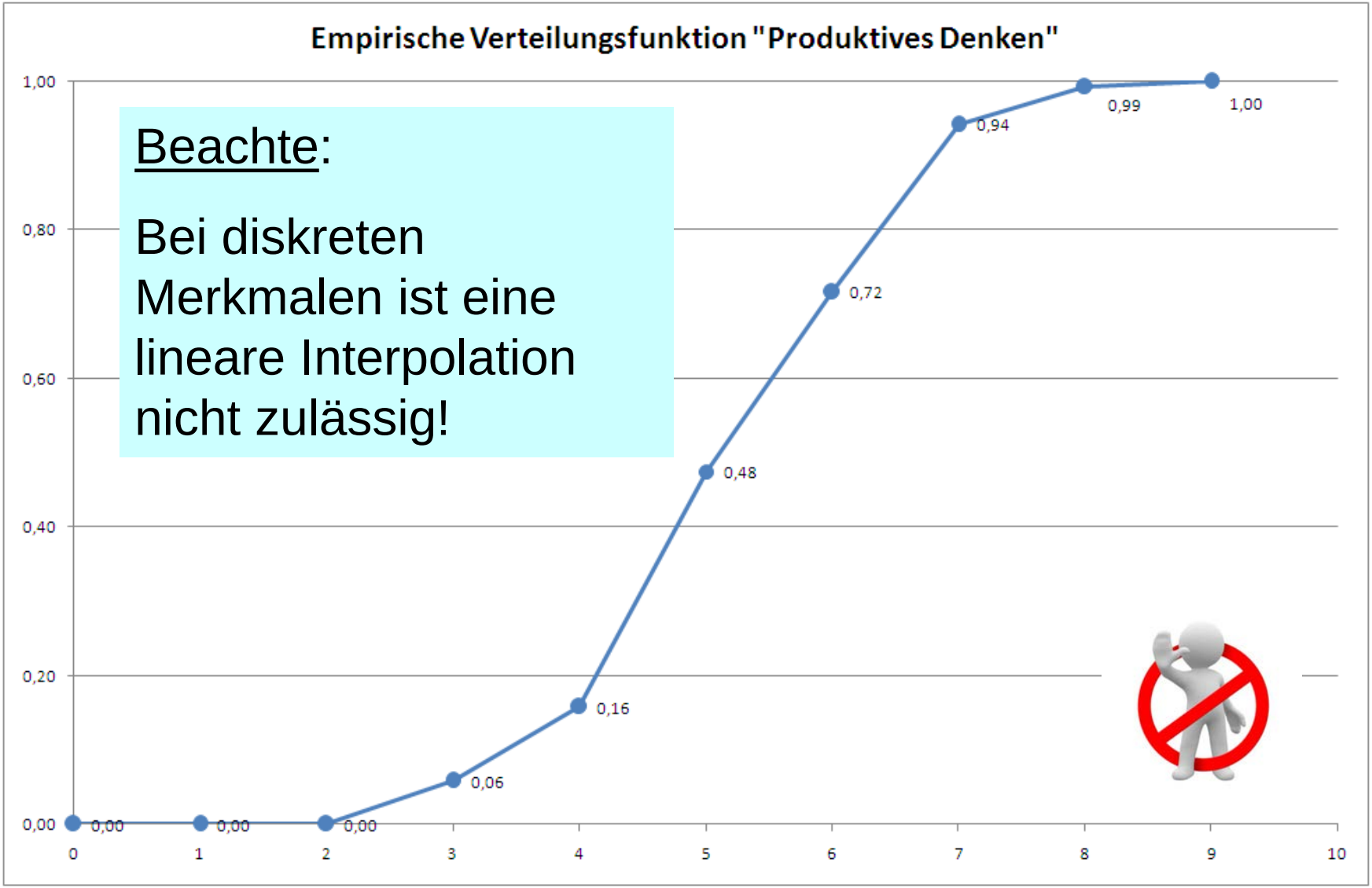
$$F(x) = H(X \leq x)$$

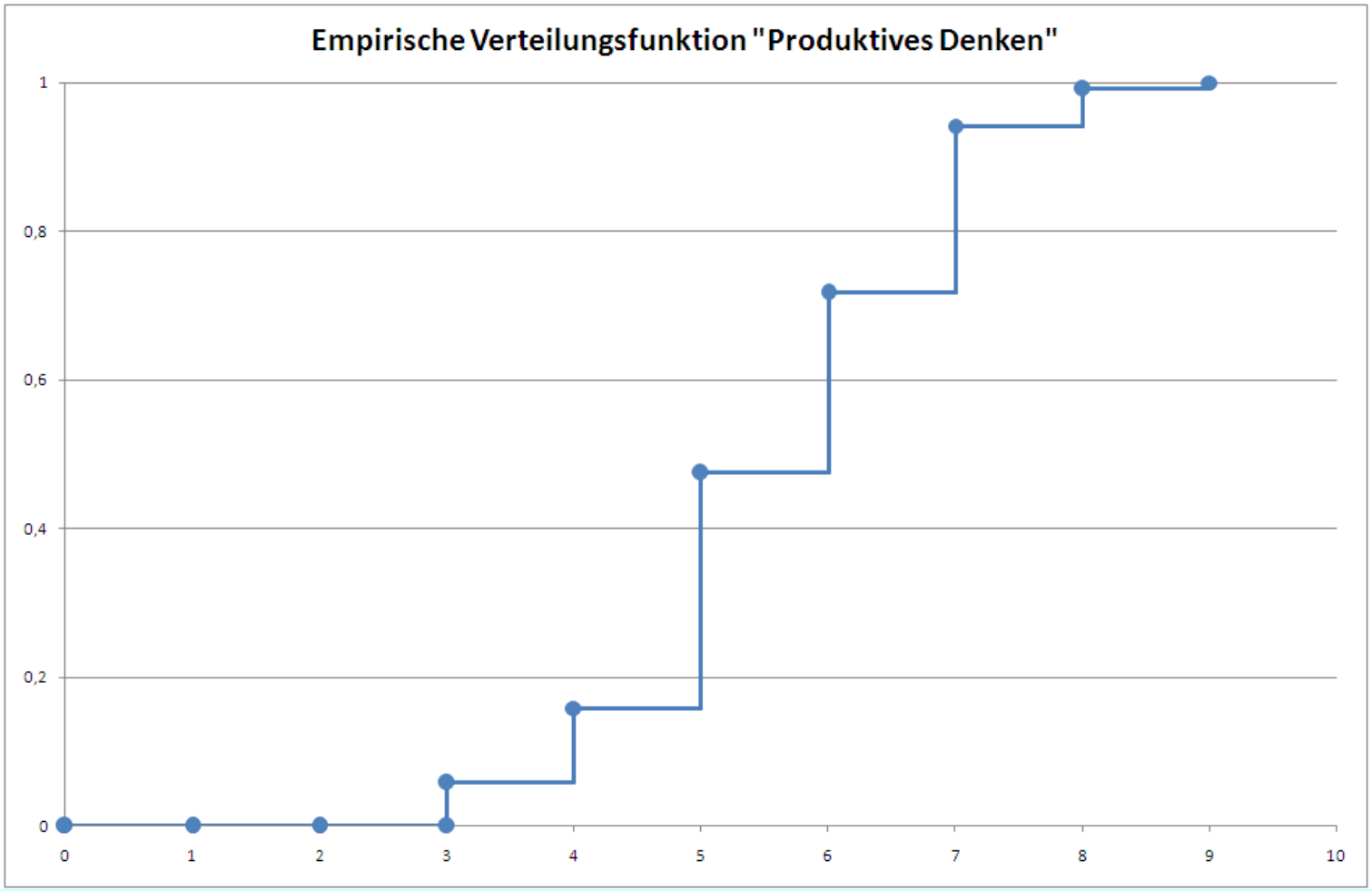
Empirische Verteilungsfunktion F(x)



72% aller Testteilnehmer haben einen Wert kleiner gleich 6 erzielt

Interpolation macht keinen Sinn

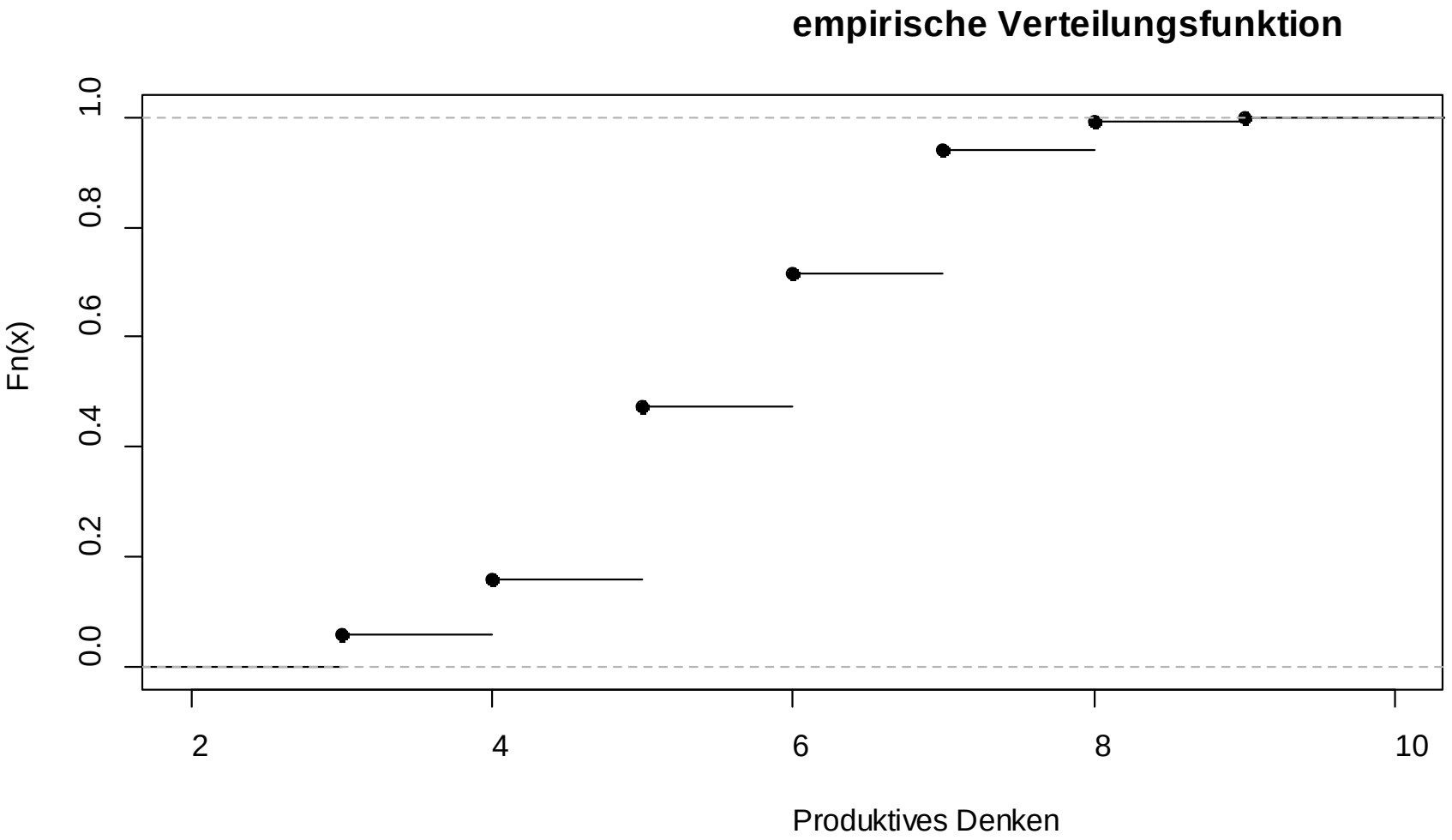




Falls ein Linienzug gewünscht wird, gibt nur eine Treppenkurve ein korrektes Bild der Verteilung

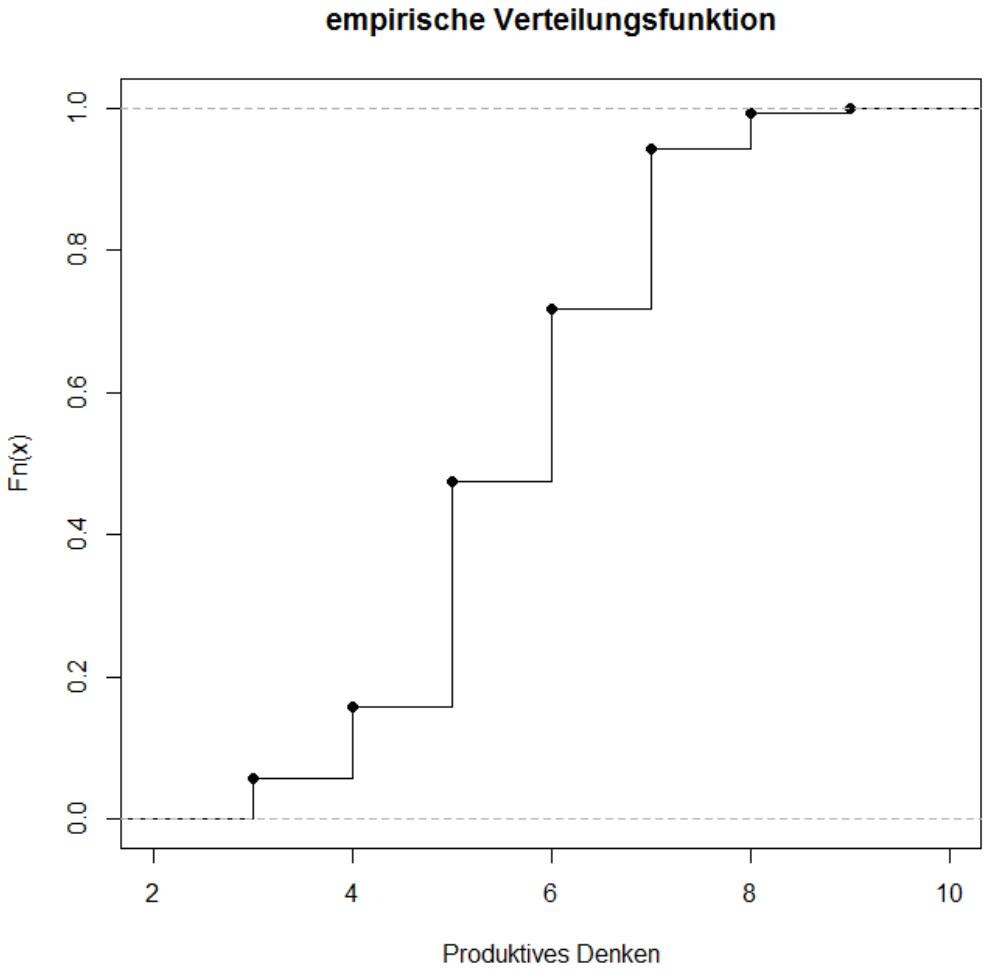
Darstellung mit R - Default

```
plot(ecdf(daten), main="empirische Verteilungsfunktion",xlab="Produktives Denken")
```



Darstellung mit R – mit Vertikalerverbindung

```
plot(ecdf(daten), main="empirische Verteilungsfunktion",
xlab="Produktives Denken", verticals=T)
```



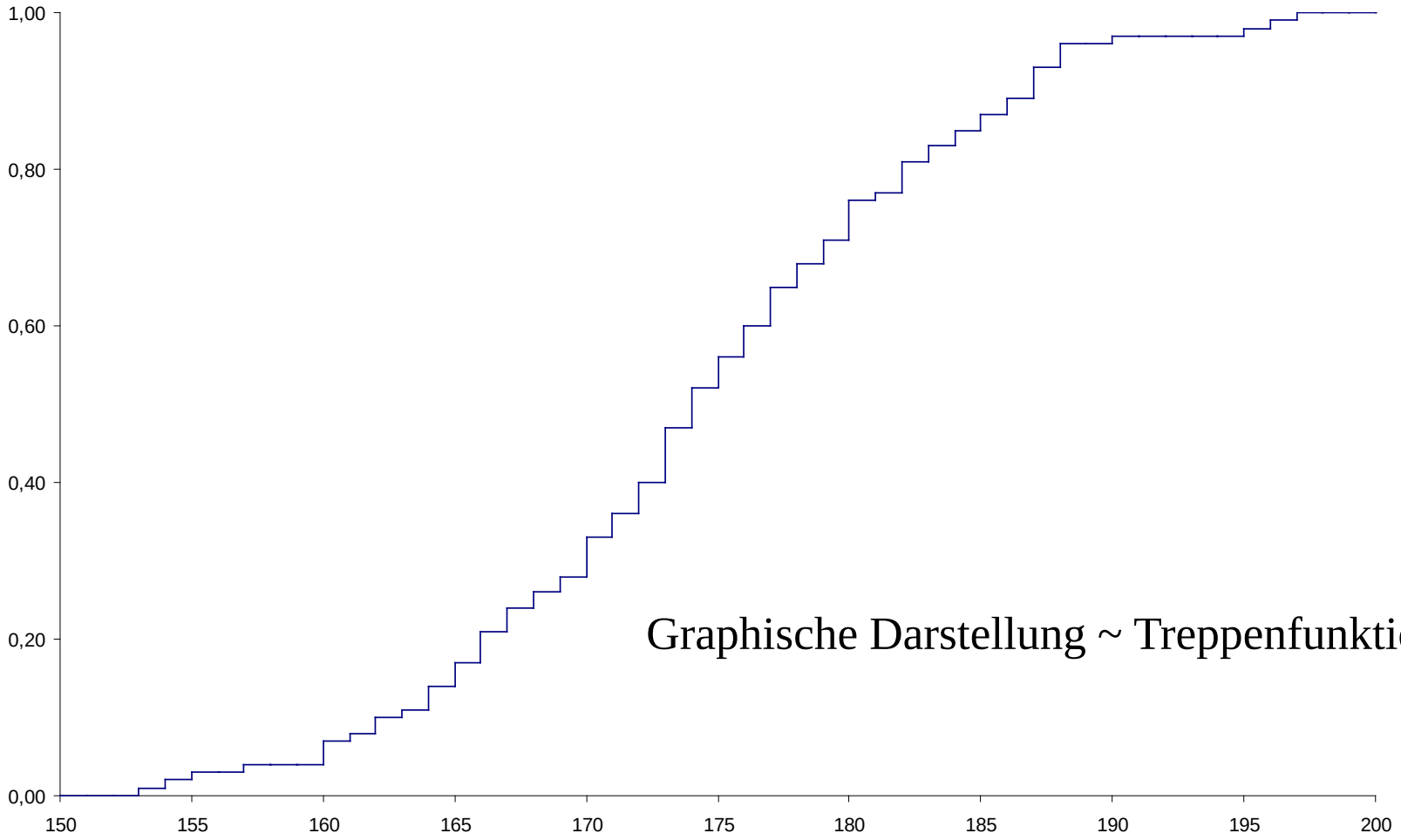
Stetiges Merkmal Kumulierte Häufigkeiten

Größe	Häufigkeit	rel. Häufigkeit	kumul. Rel. Häufigkeit
150	0	0,00	0,00
151	0	0,00	0,00
152	0	0,00	0,00
153	1	0,01	0,01
154	1	0,01	0,02
155	1	0,01	0,03
156	0	0,00	0,03
157	1	0,01	0,04
158	0	0,00	0,04
159	0	0,00	0,04
160	3	0,03	0,07
161	1	0,01	0,08
162	2	0,02	0,10
163	1	0,01	0,11
164	3	0,03	0,14
165	3	0,03	0,17
166	4	0,04	0,21
167	3	0,03	0,24
168	2	0,02	0,26
169	2	0,02	0,28
170	5	0,05	0,33
171	3	0,03	0,36
172	4	0,04	0,40
173	7	0,07	0,47
174	5	0,05	0,52
175	4	0,04	0,56

Größe	Häufigkeit	rel. Häufigkeit	kumul. Rel. Häufigkeit
176	4	0,04	0,60
177	5	0,05	0,65
178	3	0,03	0,68
179	3	0,03	0,71
180	5	0,05	0,76
181	1	0,01	0,77
182	4	0,04	0,81
183	2	0,02	0,83
184	2	0,02	0,85
185	2	0,02	0,87
186	2	0,02	0,89
187	4	0,04	0,93
188	3	0,03	0,96
189	0	0,00	0,96
190	1	0,01	0,97
191	0	0,00	0,97
192	0	0,00	0,97
193	0	0,00	0,97
194	0	0,00	0,97
195	1	0,01	0,98
196	1	0,01	0,99
197	1	0,01	1,00
198	0	0,00	1,00
199	0	0,00	1,00
200	0	0,00	1,00

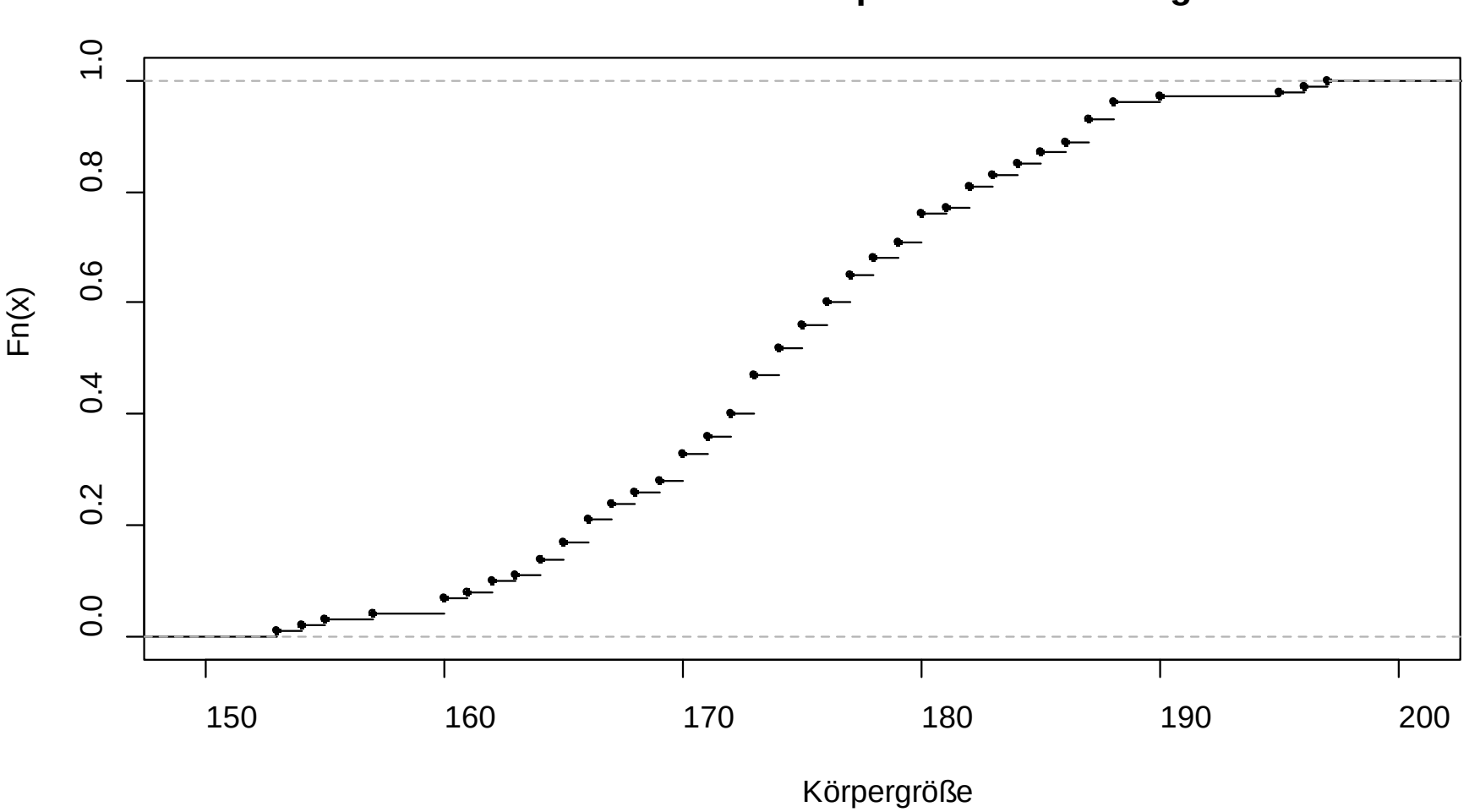
Kumulierte relative Häufigkeiten ~ Empirische Verteilungsfunktion

Empirische Verteilungsfunktion



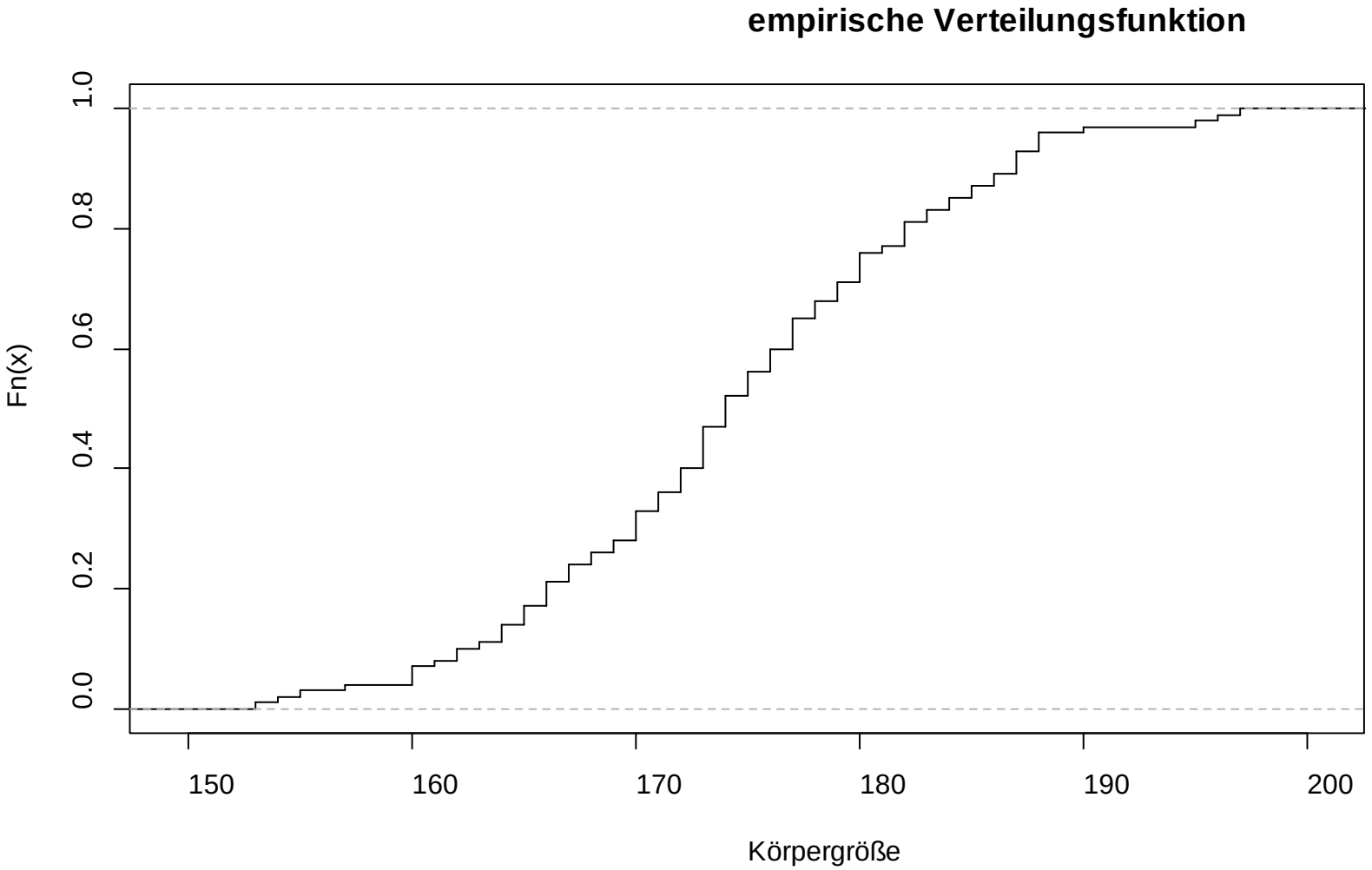
Darstellung mit R – Version 1

```
plot(ecdf(koerpergr), cex=0.5, main="empirische Verteilungsfunktion",
xlab="Körpergröße")
```

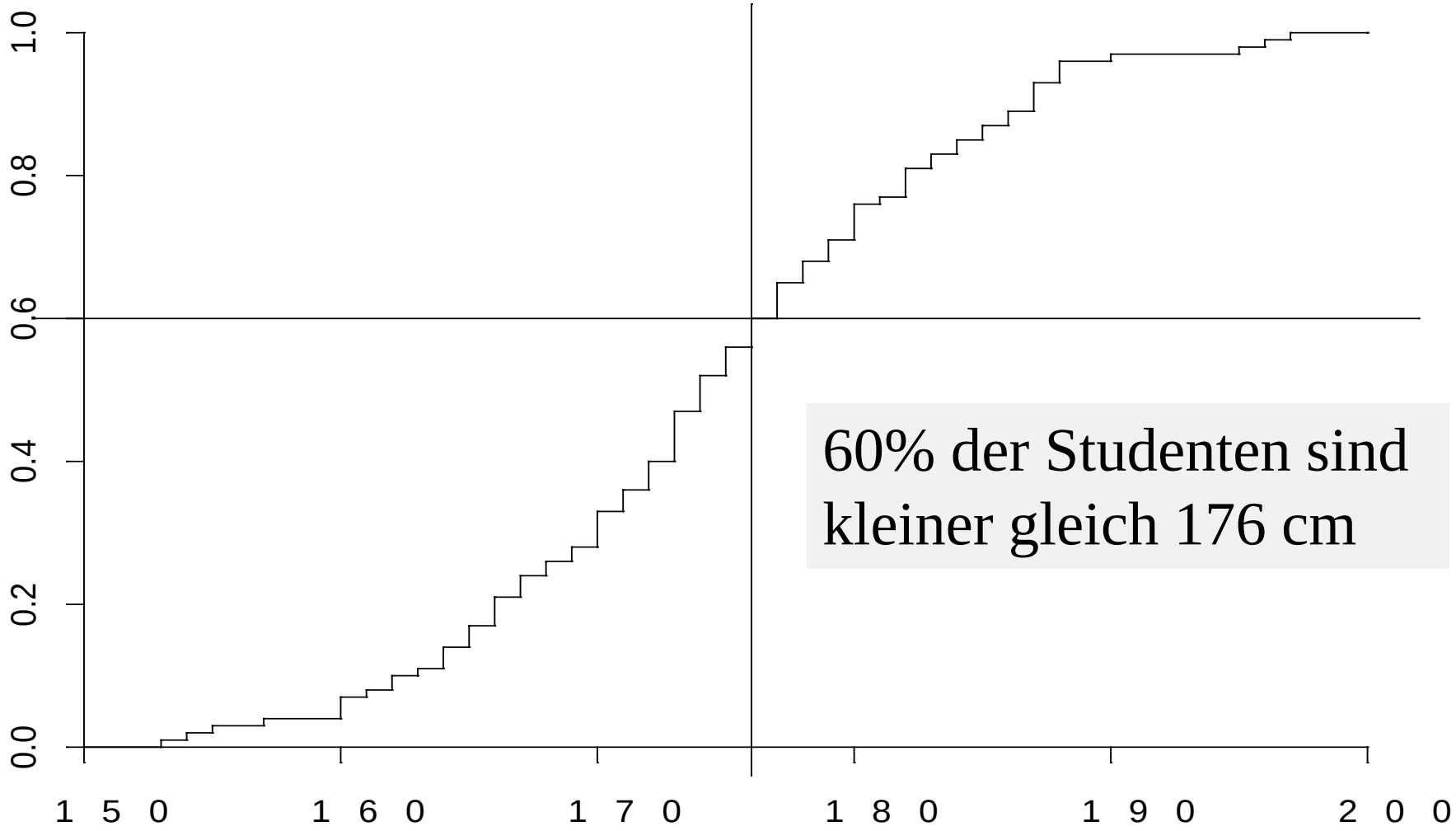


Darstellung mit R – Version 2

```
plot(ecdf(koerpergr), pch="", main="empirische Verteilungsfunktion",
verticals = T, xlab="Körpergröße")
```



Empirische Verteilungsfunktion (Leseprobe)

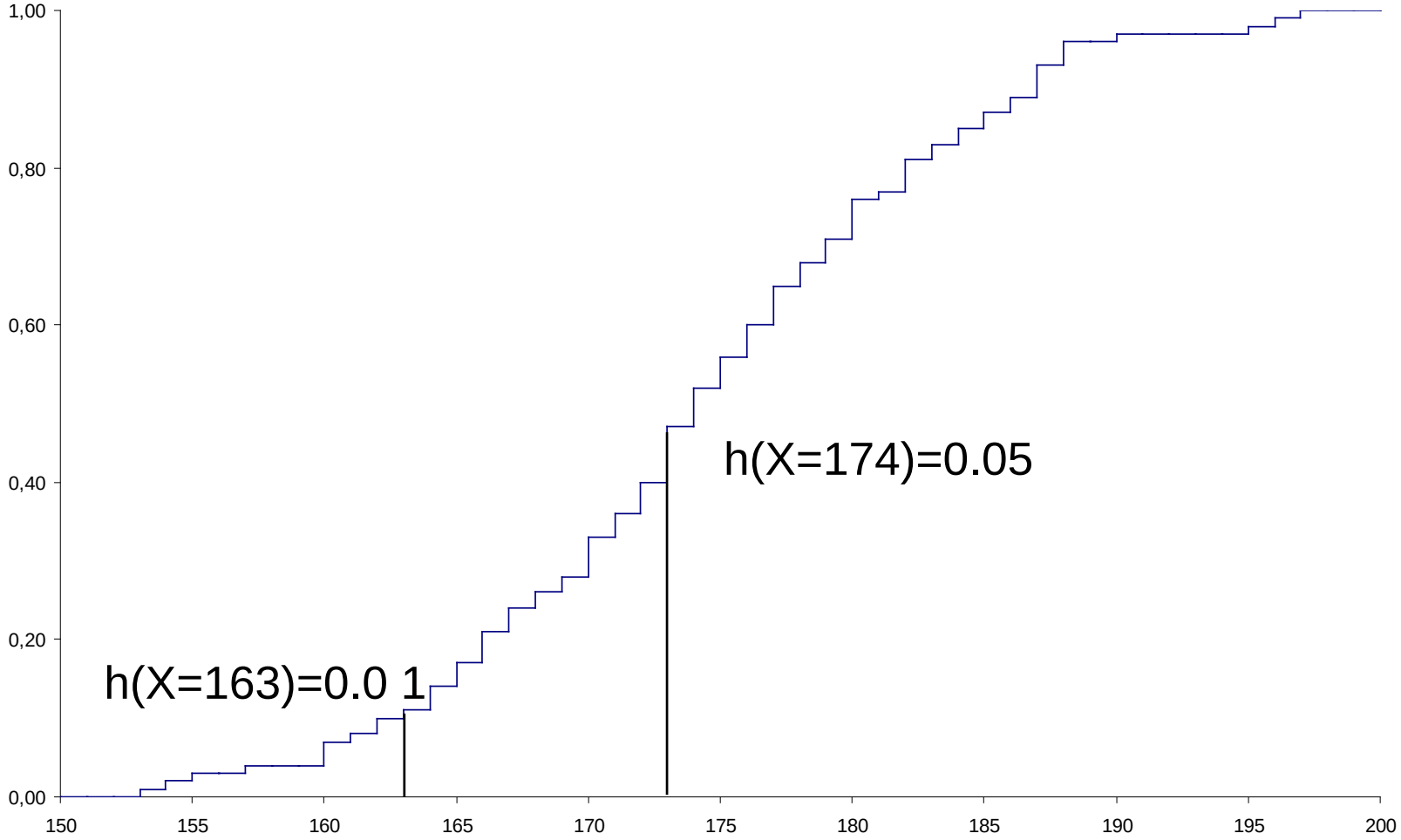


60% der Studenten sind kleiner gleich 176 cm

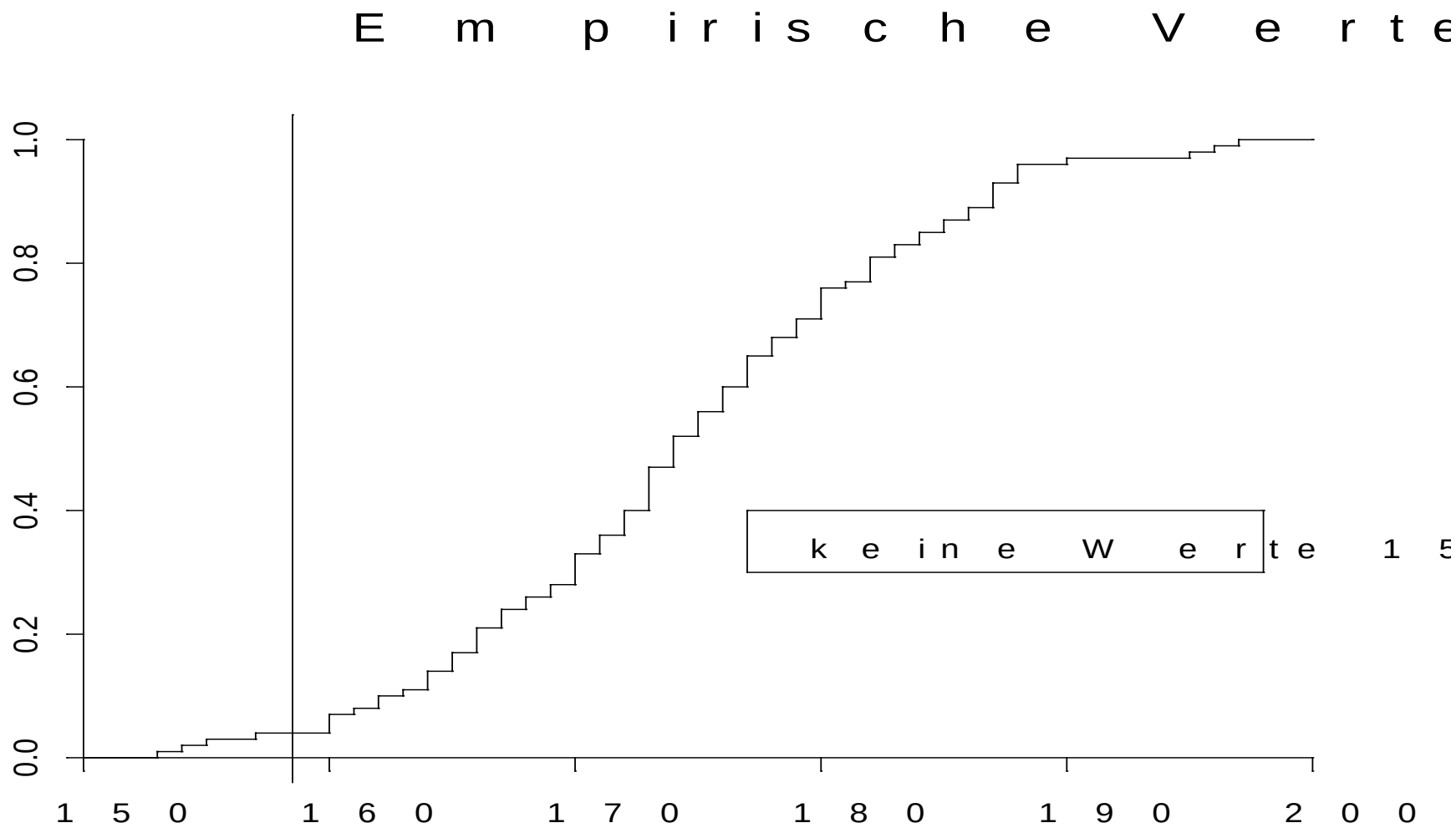
Eigenschaften der empirischen Verteilungsfunktion

- Treppenfunktion
- Bei jedem beobachteten Wert findet sich ein vertikaler Anstieg
- Die Höhe des Anstiegs beim Wert x_i ist $n(X=x_i)/n = h(x_i)$ gleich der relativen Häufigkeit dieses Wertes
- Hohe Sprünge ~ häufiger Wert
- Steiler Verlauf ~ hohe Wertedichte
- Treten in einem Wertebereich keine Werte auf, so verläuft die empirische Verteilungsfunktion in diesem Bereich horizontal

Unterschiedliche Sprunghöhen



Konstante Bereiche ~ keine Werte



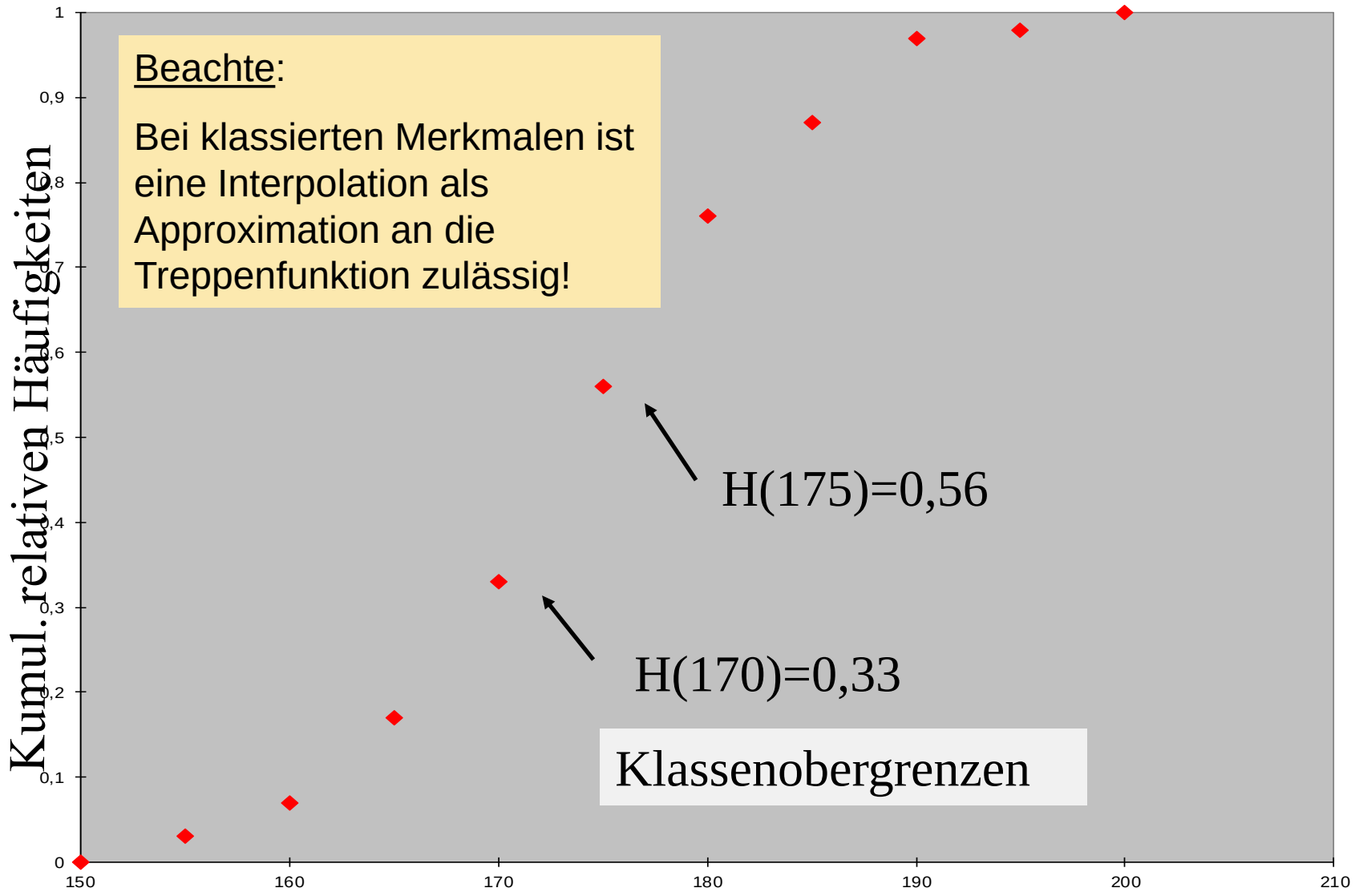
Eigenschaften der emp. Verteilungsfunktion

- Treppenfunktion
- Bei jedem beobachteten Wert findet sich ein vertikaler Anstieg
- Die Höhe des Anstiegs beim Wert x_i ist $n(X=x_i)/n = h(x_i)$
- Hohe Sprünge ~ häufiger Wert
- Steiler Verlauf ~ hohe Wertedichte
- Treten in einem Wertebereich keine Werte auf, so verläuft die emp. Verteilungsfunktion in diesem Bereich horizontal
- Die emp. Verteilungsfunktion ist monoton steigend
- Die Funktionswerte liegen zwischen 0 und 1

Kumulierte Häufigkeiten (klassierte Daten)

Bereich	n_i	h_i	N_i	H_i
150+ bis 155	3	0,03	3	0,03
155+ bis 160	4	0,04	7	0,07
160+ bis 165	10	0,10	17	0,17
165+ bis 170	16	0,16	33	0,33
170+ bis 175	23	0,23	56	0,56
175+ bis 180	20	0,20	76	0,76
180+ bis 185	11	0,11	87	0,87
185+ bis 190	10	0,10	97	0,97
190+ bis 195	1	0,01	98	0,98
195+ bis 200	2	0,02	100	1
Gesamt	100	1		

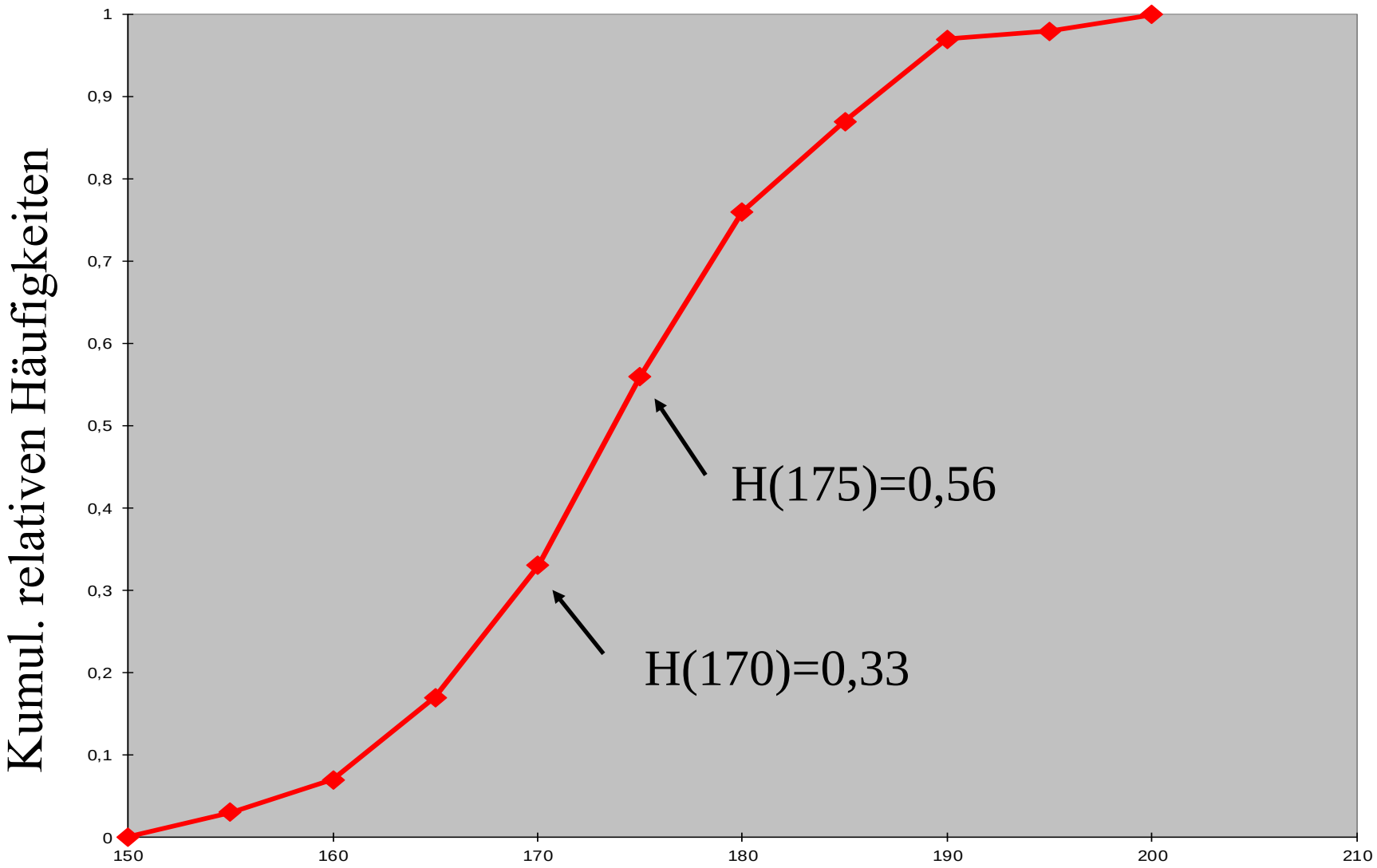
Verteilungsfunktion bei klassierten Daten



Verteilungsfunktion bei klassierten Daten

- Bei klassierten Daten können exakte Werte nur an den oberen Klassengrenzen bestimmt werden
- Ein näherungsweise Bestimmung der Werte der Verteilungsfunktion kann unter der Annahme der Gleichverteilung innerhalb der Klassen, mittels linearer Interpolation erfolgen
- In der Graphik bedeutet dies, dass wir die Punkte durch Geradenstücke zu einer durchgezogenen Linie verbinden
- Die Steigung dieser Geradenstücke entspricht der Dichte innerhalb der Klasse
- Man nennt diese Approximation der empirischen Verteilungsfunktion bei klassierten Daten auch die Summenkurve

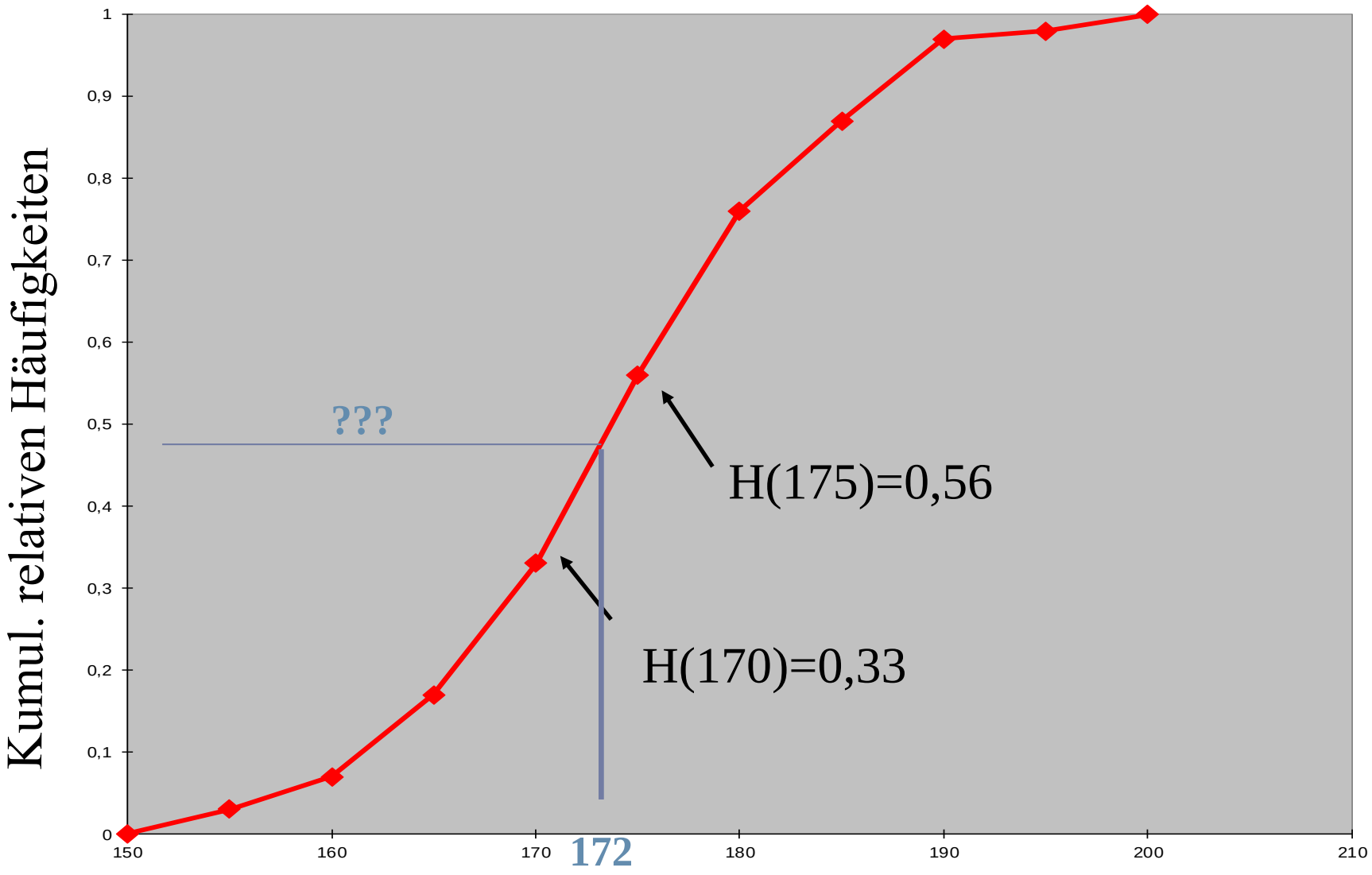
Summenkurve



Verteilungsfunktion bei klassierten Daten (Beispiel)

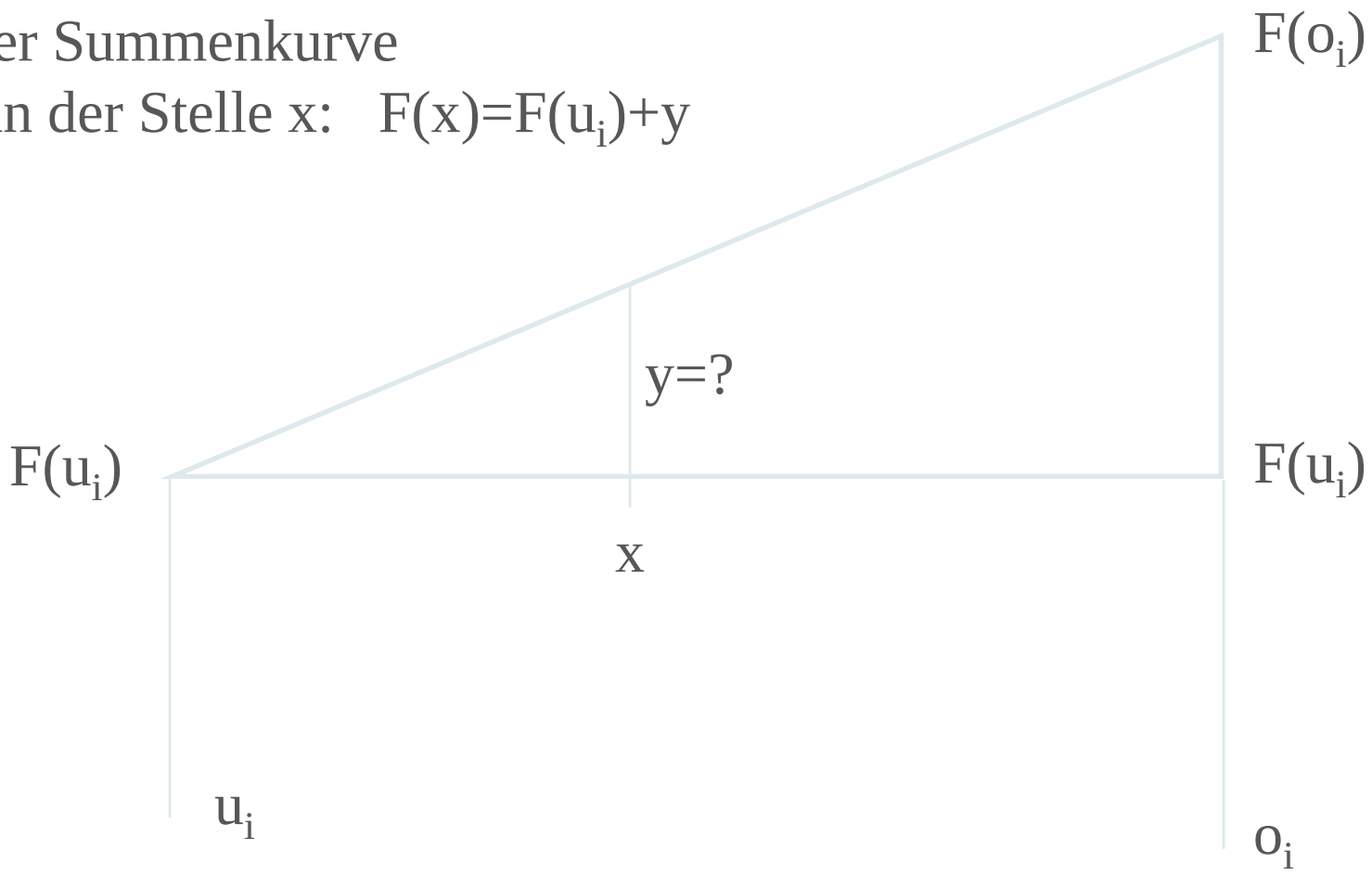
- Aus der Tabelle könne wir folgende Informationen ablesen
- 56% der Studenten sind kleiner gleich 175 cm
- 33% der Studenten sind kleiner gleich 170 cm
- Frage: Wieviel % der Studenten sind kleiner gleich 172 cm?
- Exakte Antwort aus klassierten Daten nicht mehr möglich
- Näherungsweise Lösung: Lineare Interpolation

Summenkurve



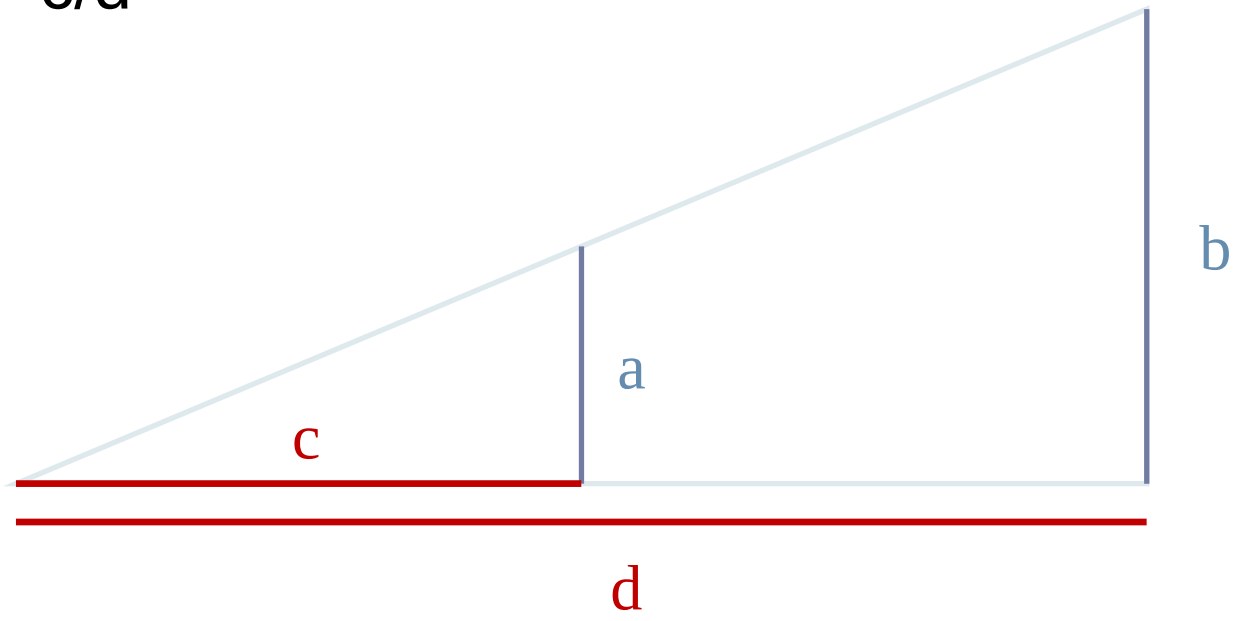
Interpolation

Gesucht ist der Funktionswert
der Summenkurve
an der Stelle x: $F(x)=F(u_i)+y$



Strahlensatz in Worten

- Das kurze vertikale Stück verhält sich zum langen vertikalen Stück genauso wie das kurze horizontale Stück zum langen horizontalen Stück
- a/b = c/d

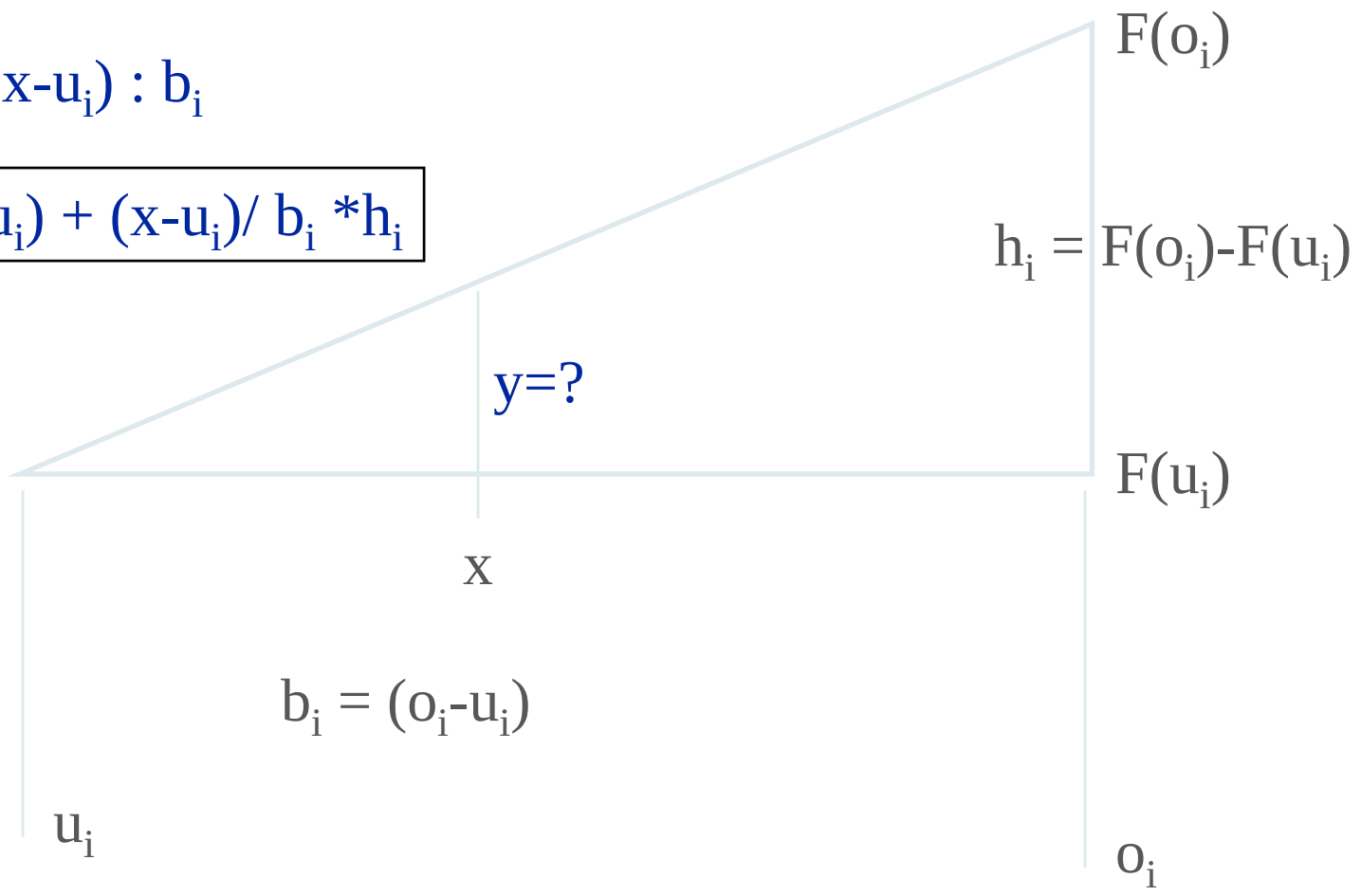


Anwendung des Strahlensatzes

$y : \{F(o_i)-F(u_i)\} = (x-u_i) : (o_i-u_i)$

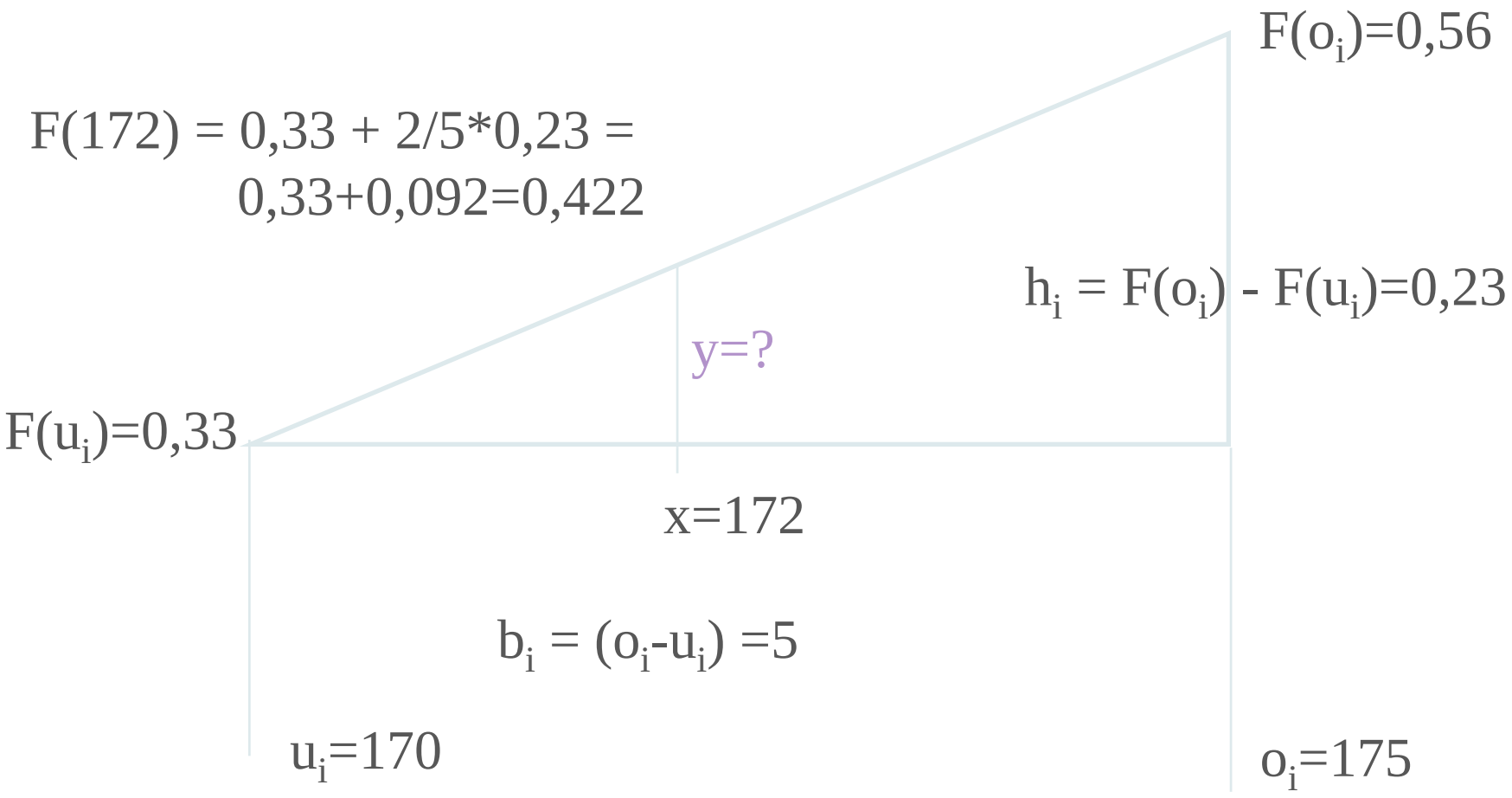
$y : h_i = (x-u_i) : b_i$

$$F(x) = F(u_i) + (x-u_i)/ b_i * h_i$$



Im Beispiel

$$F(x) = F(u_i) + (x-u_i)/ b_i * h_i$$

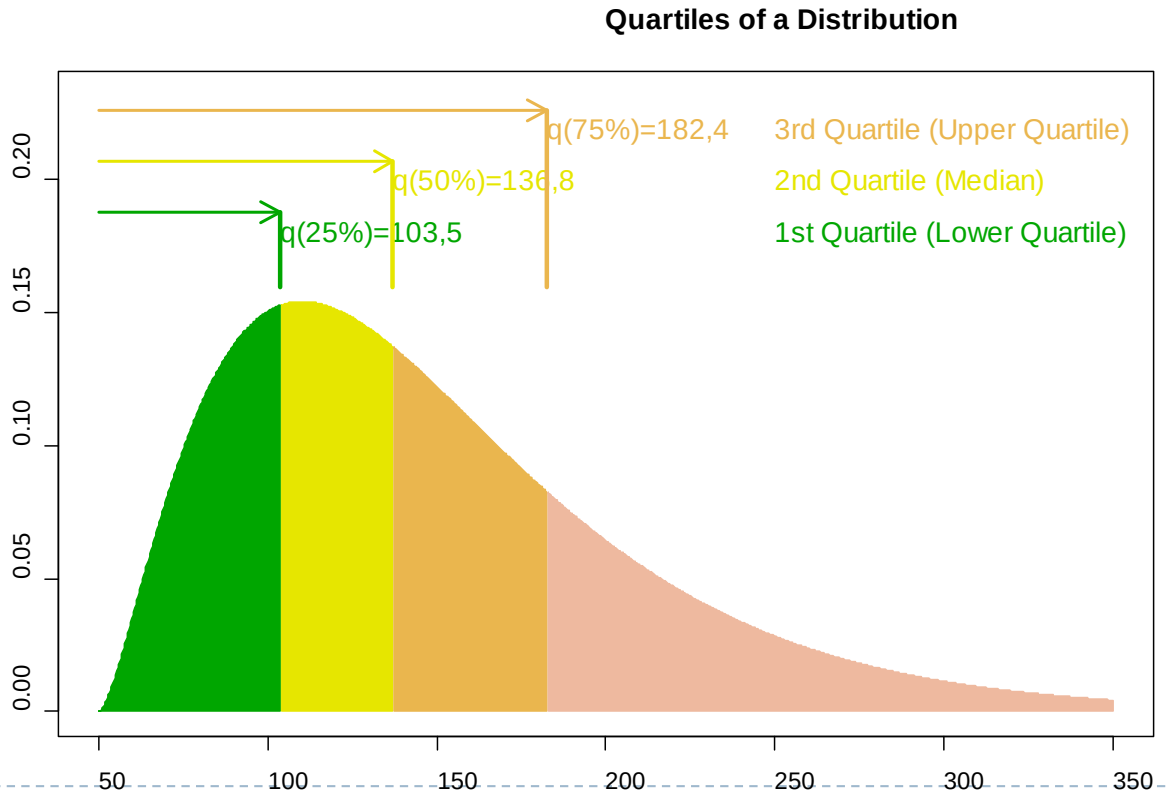


Summary

- Die Interpolation auf der Basis der Summenkurve (klassierte Daten) hat ergeben, dass 42,2% der Studenten kleiner gleich 1,72m sind.
- Auf Basis der Einzeldaten ergab sich jedoch ein Wert von 40%.
- Die Abweichung begründet sich aus dem Informationsverlust, der sich durch die Klassierung ergeben hat.
- Solche Interpolationstechniken sind für die Analyse von Sekundärdaten bedeutsam. Beachte aber dabei immer, die implizite Unschärfe.

Konzept der Quantile

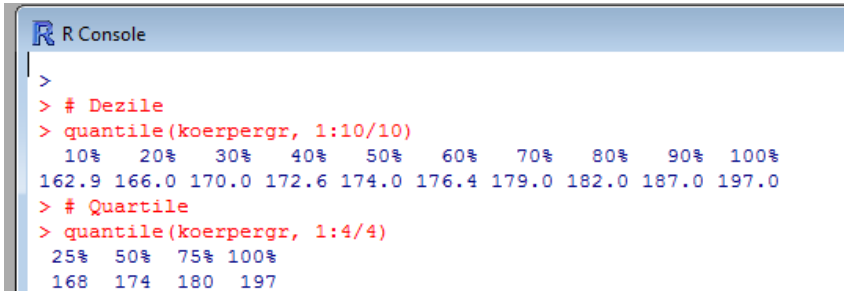
- Das Teilen eines geordneten Datensatz in q gleich große Teilmengen ist die Motivation für Quantile
- Die Quantile markieren die Grenzen zwischen aufeinanderfolgende Teilmengen



Wichtige Quantile

- **Unscharf formuliert ist ein Quantil zu einem bestimmten Prozentsatz α, jener Wert für den gilt, dass α% der Beobachtungen kleiner sind.**

- Einige wichtige Quantile, die häufig kommuniziert werden tragen einen eigenen Namen:
  - Terzile $x_{0,33}$ $x_{0,66}$
  - Quartile $x_{0,25}$ $x_{0,5}$ $x_{0,75}$
  - Dezile $x_{0,1}$ … $x_{0,9}$
  - Perzentile $x_{0,01}$, $x_{0,02}$ … $x_{0,99}$



Berechnung von Quantilen

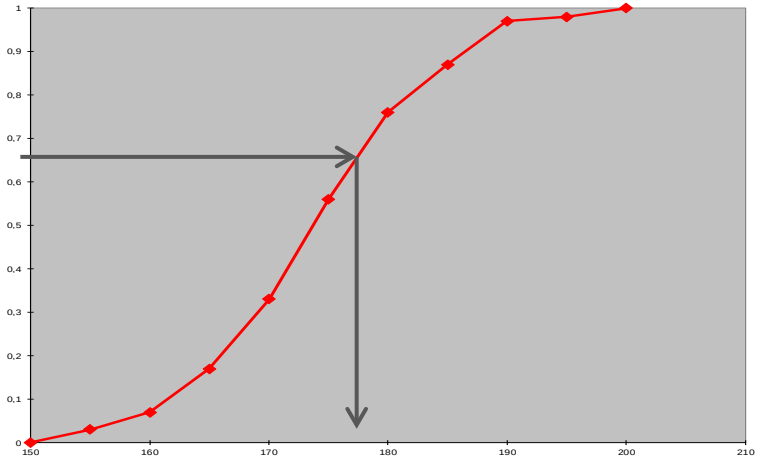
- Es gibt 2 Konzepte um Quantile zu bestimmen

  - Empirische Quantile
    Als Ergebnis wird immer ein real beobachteter Wert angegeben
  - α-Quantile

    Die Berechnung des Quantils erfolgt mittels Interpolation

Empirisches Quantil

- Ausgehend von einem Anteilswert p (y-Achse) wird der zugehörige Wert aus der Stichprobe bestimmt, für den F(x) zum ersten mal größer als oder zumindest gleich groß wie p ist.
- Das bedeutet ein empirisches p-Quantil ist jener möglichst kleine Merkmalswert für den gerade noch gilt, dass p-Prozent der Beobachtungen kleiner gleich als eben dieser Merkmalswert sind.



Merke:
Empirische Quantile sind immer nur real beobachtete Werte.

Definition: Empirisches Quantil

- $0 < p < 1$
- Datensatz: $x_1, ..., x_n$
- Das Empirische p-Quantil x_p

  ist dann der kleinste beobachtete Wert x für

  den bereits gilt: $F(x) \geq p$
- Seien $x_{(1)}, ... x_{(n)}$ die geordneten Werte:
- $x_p = x_{(k),}$ ist dann der k-te Wert in der geordneten Stichprobe, wobei k wie folgt bestimmt wird:

  $$(k-1)/n < p \leq k/n \quad \text{bzw.} \quad (k-1) < np \leq k$$

Mini-Beispiel zu empirischen Quantilen

- Stichprobe: 3, 6, 2, 8, 7, 5, 9, 4 n=8
- Geordnete Stichprobe: 2,3,4,5,6,7,8,9
- Wir wollen das as Empirische p-Quantil für p=0,25 bestimmen
- $x_{0,25} = ?$
- $(k-1)/n < p \leq k/n$ bzw. $(k-1) < np \leq k$
- $(k-1)/8 < 0,25 \leq k/8$ bzw. $(k-1) < 2 \leq k$ ➔ k=2
- $x_{0,25} = x_{(2)} = 3$
- $x_{0,75} = ?$
- $(k-1)/8 < 0,75 \leq k/8$ bzw. $(k-1) < 6 \leq k$ ➔ k=6
- $x_{0,75} = x_{(6)} = 7$

Weitere Beispiele zu empirischen Quantilen

- Gesucht ist ein Wert, so dass 95% der Studenten kleiner gleich diesem Wert sind
- Datensatz Körpergröße n=100 p=0,95

  $x_{0,95}$ = ?

  $(k-1)/n < p \leq k/n \Rightarrow (k-1) < np \leq k$

  $(k-1) < 95 \leq k \Rightarrow k=95$

  $x_{0,95}$ = 188
- Datensatz produktives Denken n=120

  $x_{0,50}$ = ? p=0,5

  $(k-1) < 120*0,5 \leq k \Rightarrow k=60$

  $x_{0,50}$ = 7

α-Quantile

- Wenn wir uns nicht auf reale Beobachtungen beschränken, sondern auch Interpolationen zwischen beobachteten Werten zulassen, kommen wir zu den sog. α-Quantilen
- Seien $x_{(1)}, \ldots, x_{(n)}$ die geordneten Beobachtungen einer Urliste, so ist das α-Quantil wie folgt definiert:

$$\tilde{x}_\alpha = \begin{cases} x_{(k)} & k = \lceil n \cdot \alpha \rceil \\ \frac{1}{2}(x_{(k)} + x_{(k+1)}) & k = n \cdot \alpha \end{cases}$$

Falls $n.\alpha$ keine ganze Zahl ist: Kleinste ganze Zahl größer gleich

Falls $n.\alpha$ eine ganze Zahl ist;

α-Quantile

- α-Quantile sind so definiert, dass n.α Beobachtungswerte kleiner und n.(1-α) Beobachtungswerte größer als das jeweilige α-Quantil sind.
- Falls n.α eine ganze Zahl ist, würde im Prinzip jeder beliebige Wert zwischen $x_{(n.\alpha)}$ und $x_{(n.\alpha+1)}$ die obige Bedingung erfüllen.
- Die von uns angegebene Formel nimmt einfach die Mitte zwischen $x_{(n.\alpha)}$ und $x_{(n.\alpha+1)}$
- Beachte: Unterschiedliche Softwaresysteme verwenden leicht unterschiedliche Definitionen von α-Quantilen, was insbesondere bei kleinen Stichproben deutliche Unterschiede ausmachen kann

Beispiel α-Quantile

- Urliste: x_1 , x_2, x_3, x_4, …, x_{10}
  8, 5, 7, 14 , 27 12, 24, 17, 3, 21
- Geordnete Urliste: $x_{(1)}$, $x_{(2)}$, $x_{(3)}$, $x_{(4)}$, …, $x_{(10)}$
  3, 5, 7, 8, 12, 14, 17, 21, 24, 27

- 1.Quartil; 25%-Quantil; Q(25%)
  - *α=0,25 n*α=2.5 ➔ Index=3*
  - Q(25%)=7

$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$	$x_{(6)}$	$x_{(7)}$	$x_{(8)}$	$x_{(9)}$	$x_{(10)}$
3	5	7	8	12	14	17	21	24	27

Beispiel α-Quantile

```
> x <- c(8, 5, 7, 14 , 27,  12, 24, 17, 3, 21)
> x
 [1]  8  5  7 14 27 12 24 17  3 21
> quantile(x, 0.3)
30%
7.7
>
```

- Urliste: x_1 , x_2, x_3, x_4, …, x_{10}
  8, 5, 7, 14 , 27 12, 24, 17, 3, 21
- Geordnete Urliste: $x_{(1)}$, $x_{(2)}$, $x_{(3)}$, $x_{(4)}$, …, $x_{(10)}$
  3, 5, 7, 8, 12, 14, 17, 21, 24, 27

- 3.Dezil; 30%-Quantil; Q(30%)
  - *α=0,30 n*α = 3 ➔ Index=3 oder 4*
  - Unsere Formel:
  - Q(30%)=(7+8)/2=7.5

$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$	$x_{(6)}$	$x_{(7)}$	$x_{(8)}$	$x_{(9)}$	$x_{(10)}$
3	5	7	8	12	14	17	21	24	27

Excel-Funktion Quantil (auch quantile von R) liefert 7,7

SPSS-Funktion Frequencies liefert 7,3

Anwendung von Dezilen

Männl. Angestellte ⟷ Männl. Beamte

Statistischer Annex	Bruttojahreseinkommen 2005 (in €), Männer		
	Angestellte	**Beamte**	**Angestellte -/+ ...% als Beamte**
Median	36.138	41.839	-13,6
Arithmetisches Mittel	42.428	46.370	-8,5
6. Dezil	42.000	45.409	-7,5
7. Dezil	49.307	50.551	-2,5
8. Dezil	59.120	57.567	+2,7
9. Dezil	76.793	70.822	+8,4

Martin Bauer/Julia Schuster AEB 2006, 22.3.2007 16

Quelle:
Statistik Austria

Quantile bei klassierten Daten

- Bei klassierten Daten ergibt sich das α-Quantil durch Interpolation
- Ausgangspunkt ist jene Klasse, in der die kumulierten Häufigkeiten den Wert α übersteigen
- Zur Berechnung verwenden wir, wie zuvor den Strahlensatz, allerding sind wir nun an der Bestimmung des kurzen horizontalen Stücks interessiert.
- Zunächst muss immer die relevante Klasse gefunden werden

Bestimmung des 0,5 Quantils

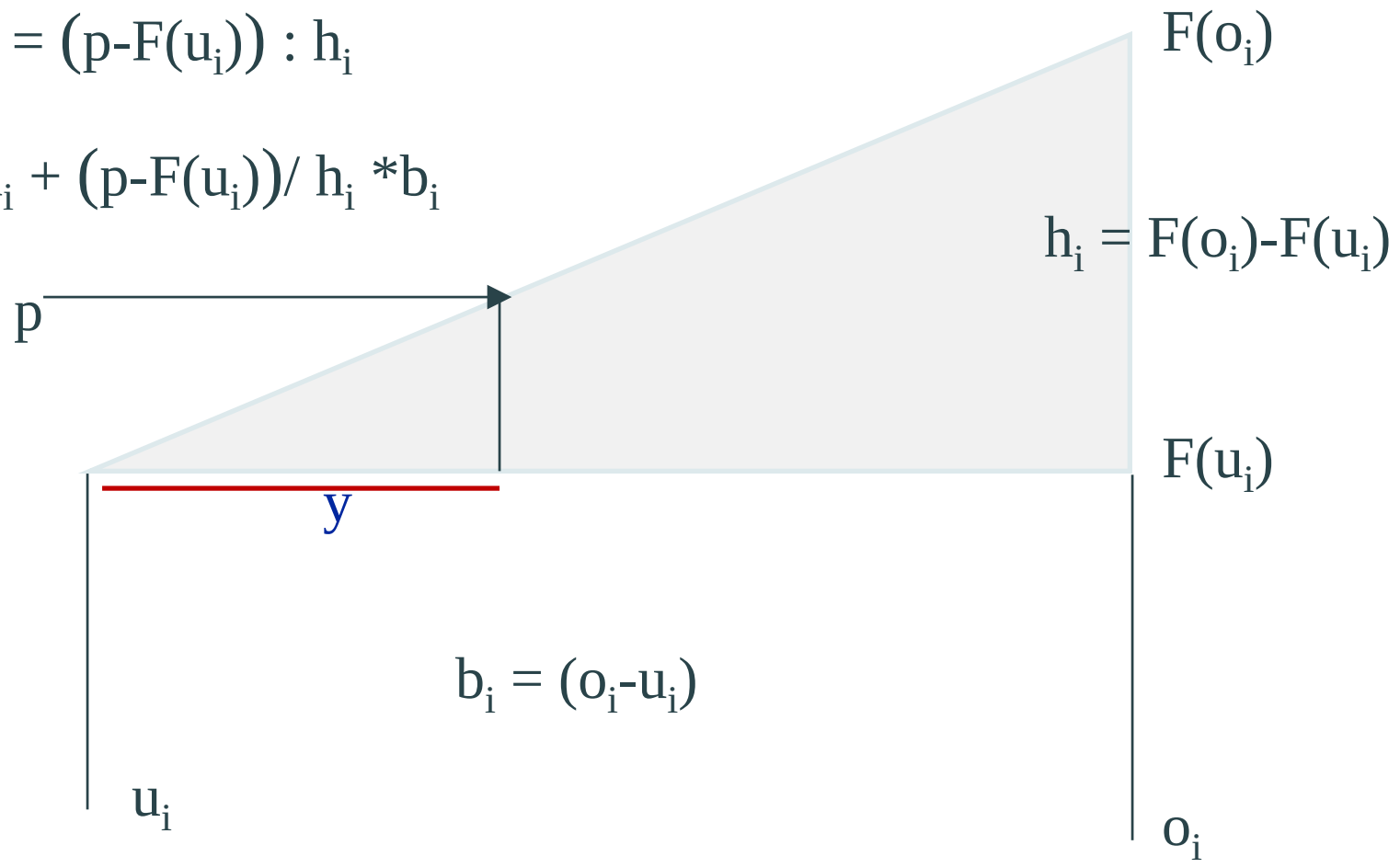
Bereich	n_i	h_i	N_i	H_i
150+ bis 155	3	0,03	3	0,03
155+ bis 160	4	0,04	7	0,07
160+ bis 165	10	0,10	17	0,17
165+ bis 170	16	0,16	33	0,33
170+ bis 175	23	0,23	56	0,56
175+ bis 180	20	0,20	76	0,76
180+ bis 185	11	0,11	87	0,87
185+ bis 190	10	0,10	97	0,97
190+ bis 195	1	0,01	98	0,98
195+ bis 200	2	0,02	100	1
Gesamt	100	1		

Wo überschreiten die kumulierten Häufigkeiten den vorgegebenen Prozentwert?

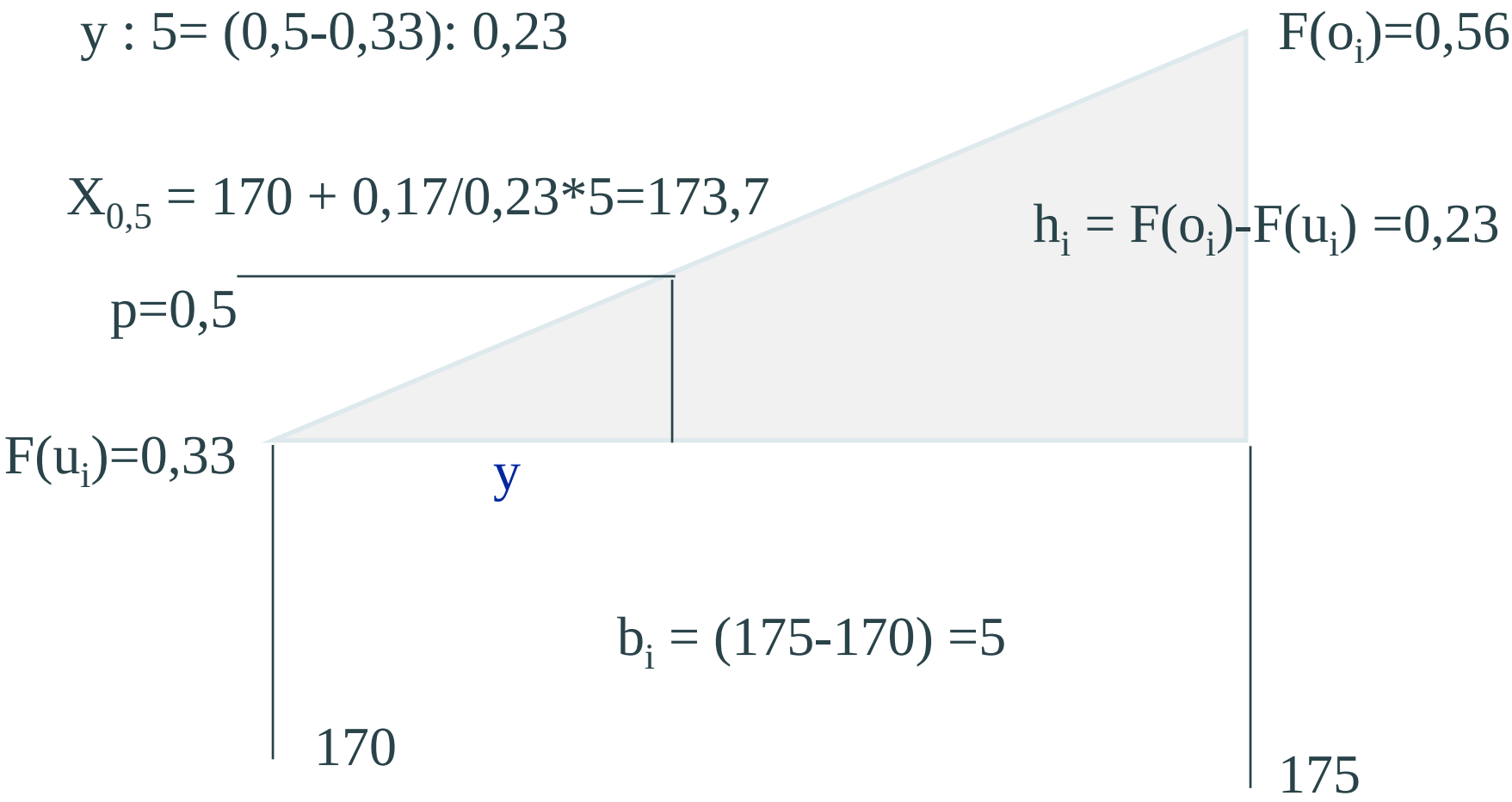
Quantile bei klassierten Daten

$$y : b_i = (p\text{-}F(u_i)) : h_i$$

$$x_p = u_i + (p\text{-}F(u_i))/\ h_i * b_i$$



Quantile bei klassierten Daten



Quantile

Beispiel: Körpergröße (Originalwerte)

1.Quartil = $x_{0.25}$
2.Quartil = $x_{0.50}$
3.Quartil = $x_{0.75}$

Five Number Summary

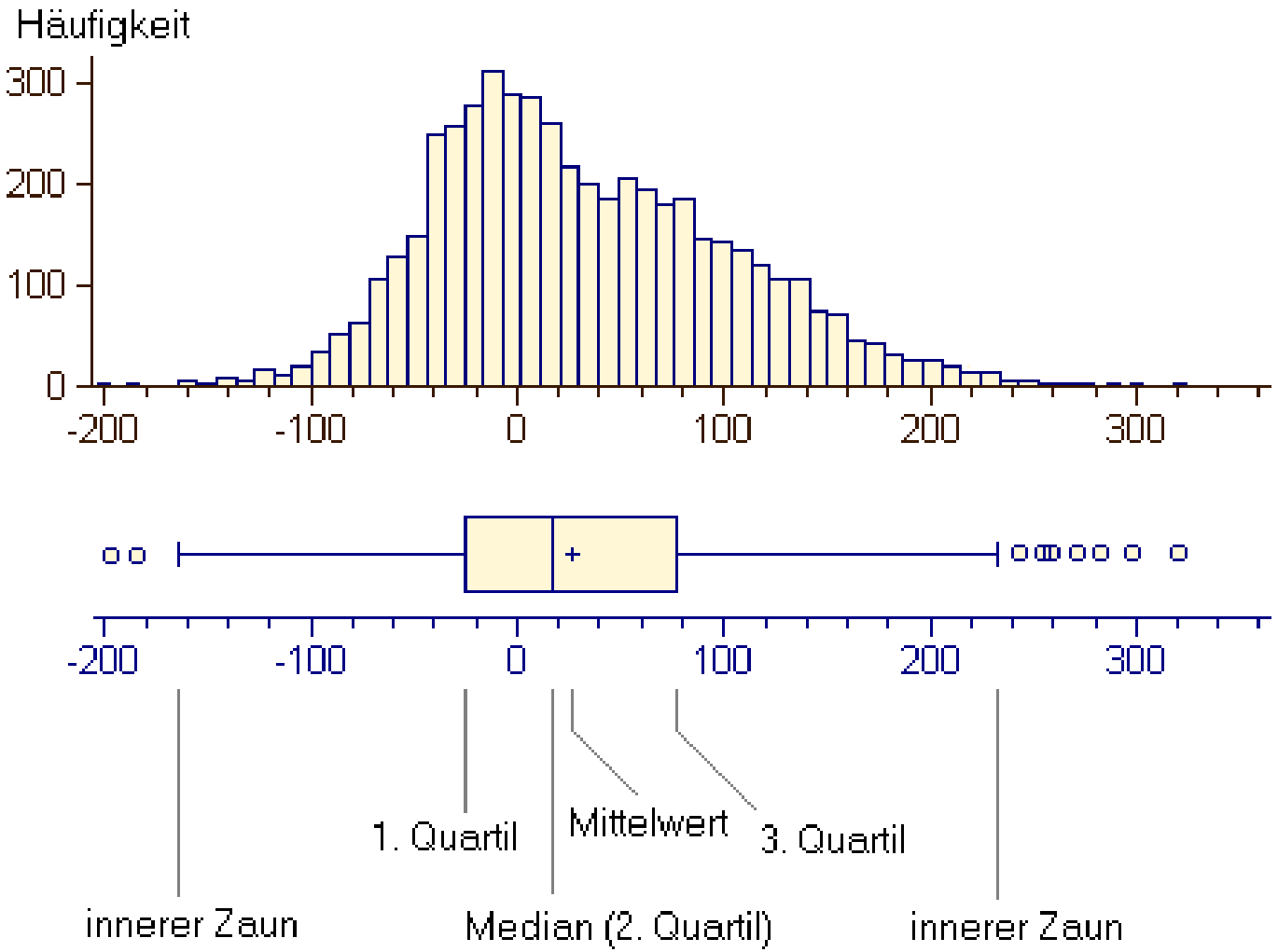
Min.	1st Qu.	2nd Qu.	3rd Qu.	Max.
153	168	174	180	197
$x_{(1)}$	$x_{(25)}$	$x_{(50)}$	$x_{(75)}$	$x_{(100)}$

Box-Plots

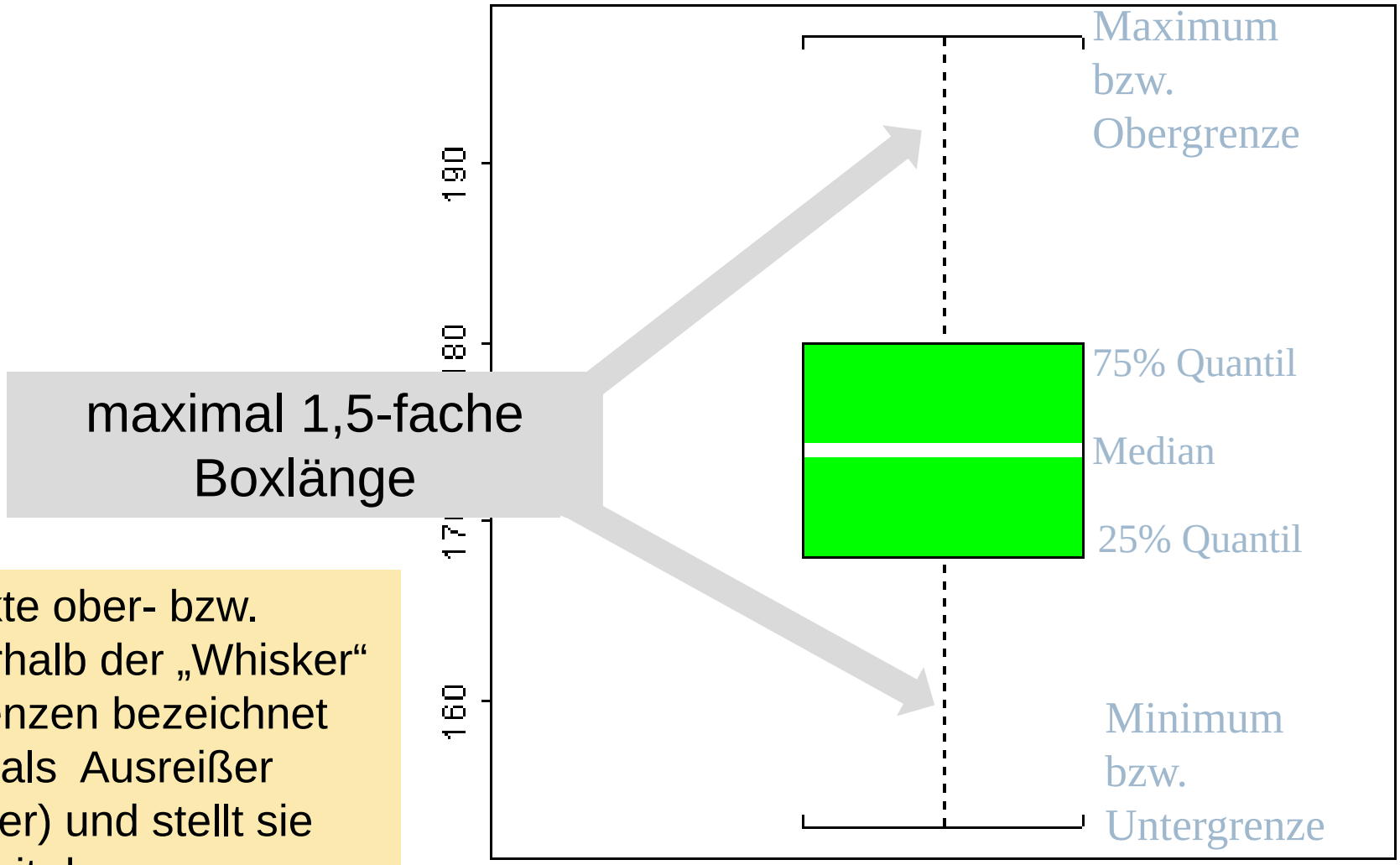
- Basierend auf den 5 zusammenfassenden Werten einer Verteilung:
  - Minimum, 1.Quartil, 2.Quartil, 3.Quartil und Maximum

  lassen sich instruktive Graphiken zur Darstellung einer Verteilung entwickeln, die insbesondere zum Vergleich mehrerer Gruppen gut geeignet sind.
- Häufig werden die begrenzenden Linien nicht bis zum Minimum und Maximum der Daten gezogen. Die Balkenlänge wird mit der 1,5-fachen Boxhöhe begrenzt und extreme Datenwerte werden extra markiert.

Boxplot

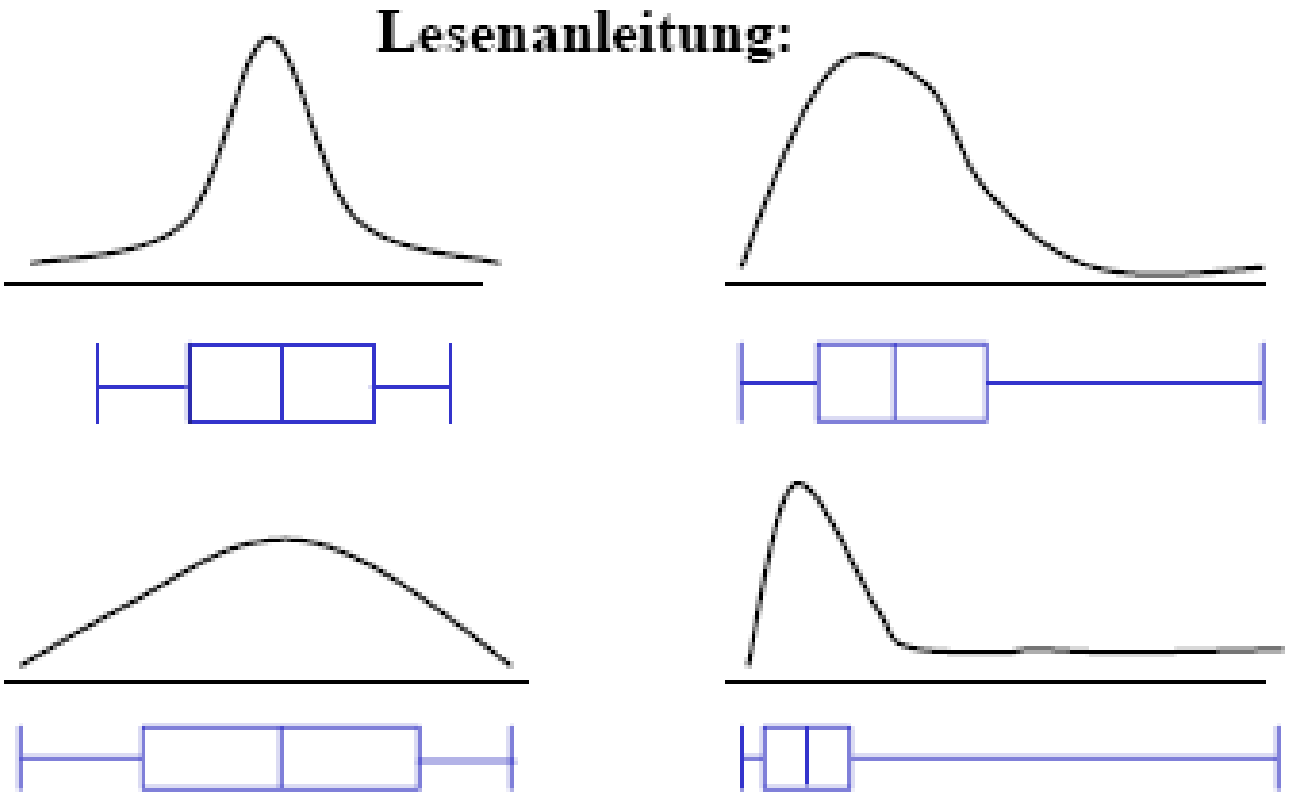

Häufigkeit
300
200
100
0
-200
-100
0
100
200
300
-200
-100
0
100
200
300
1. Quartil
Mittelwert
3. Quartil
innerer Zaun
Median (2. Quartil)
innerer Zaun

Boxplot (Box-Whisker-Plot)

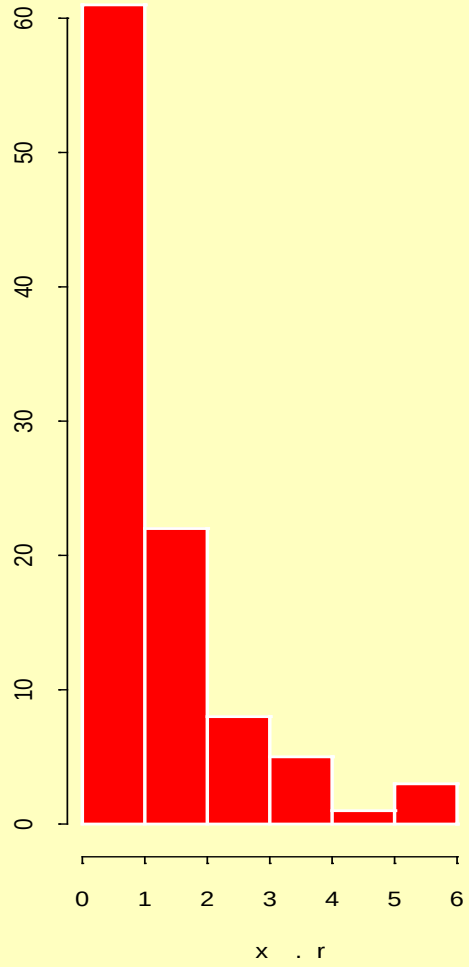


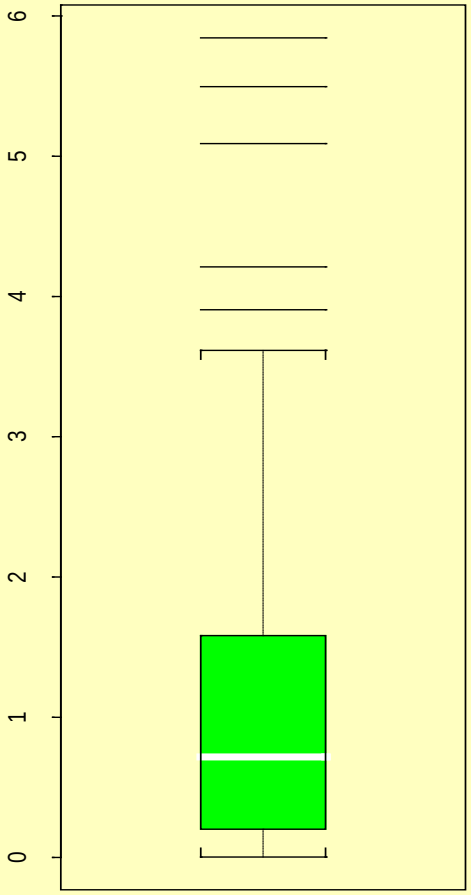
Punkte ober- bzw. unterhalb der „Whisker“ - Grenzen bezeichnet man als Ausreißer (outlier) und stellt sie explizit dar

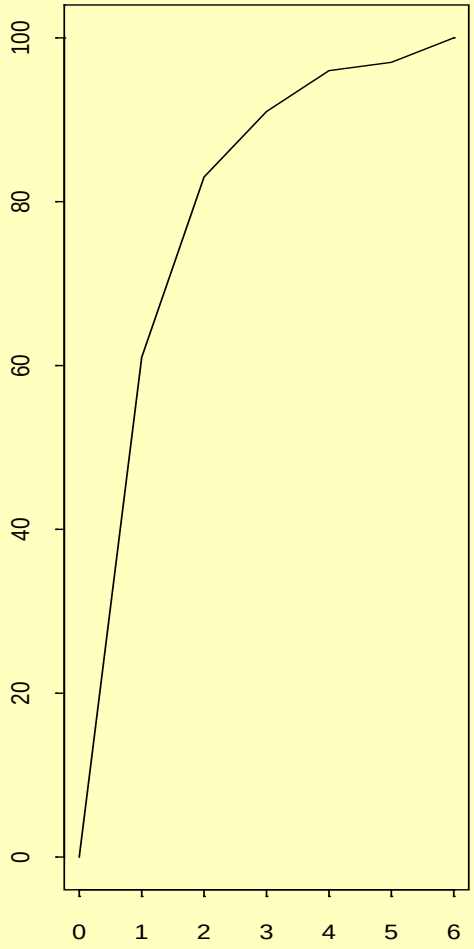
Zur Interpretation von Boxplots



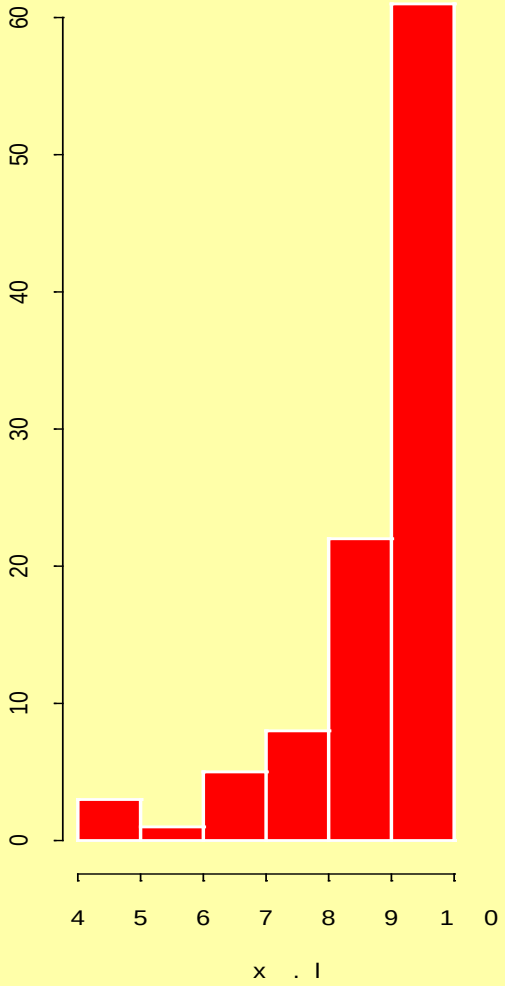
Beispiel einer rechtsschiefen Verteilung

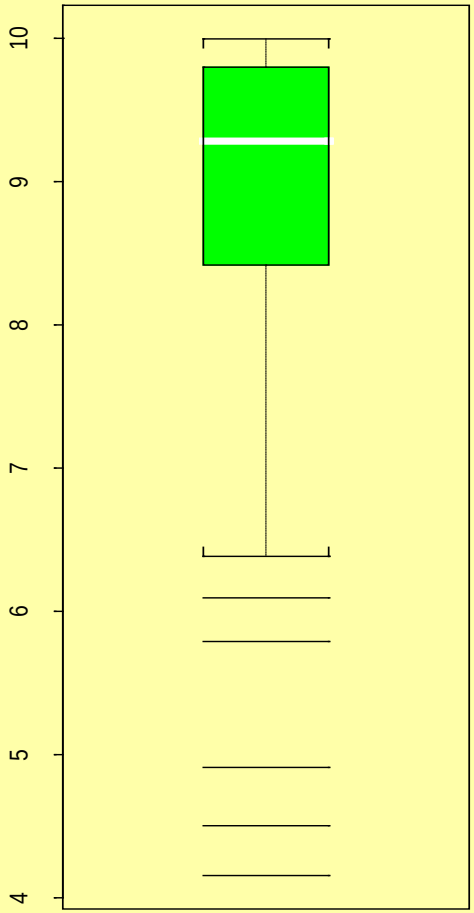


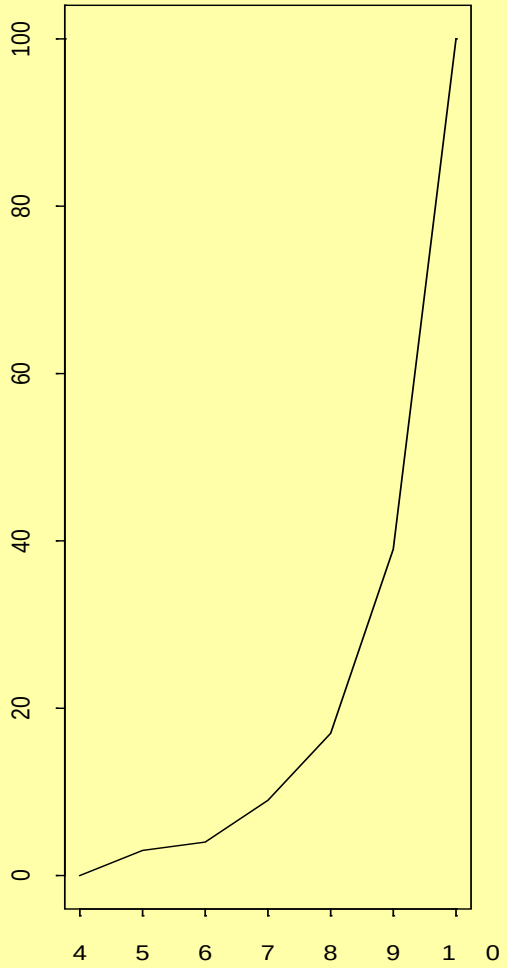




Beispiel einer linksschiefen Verteilung



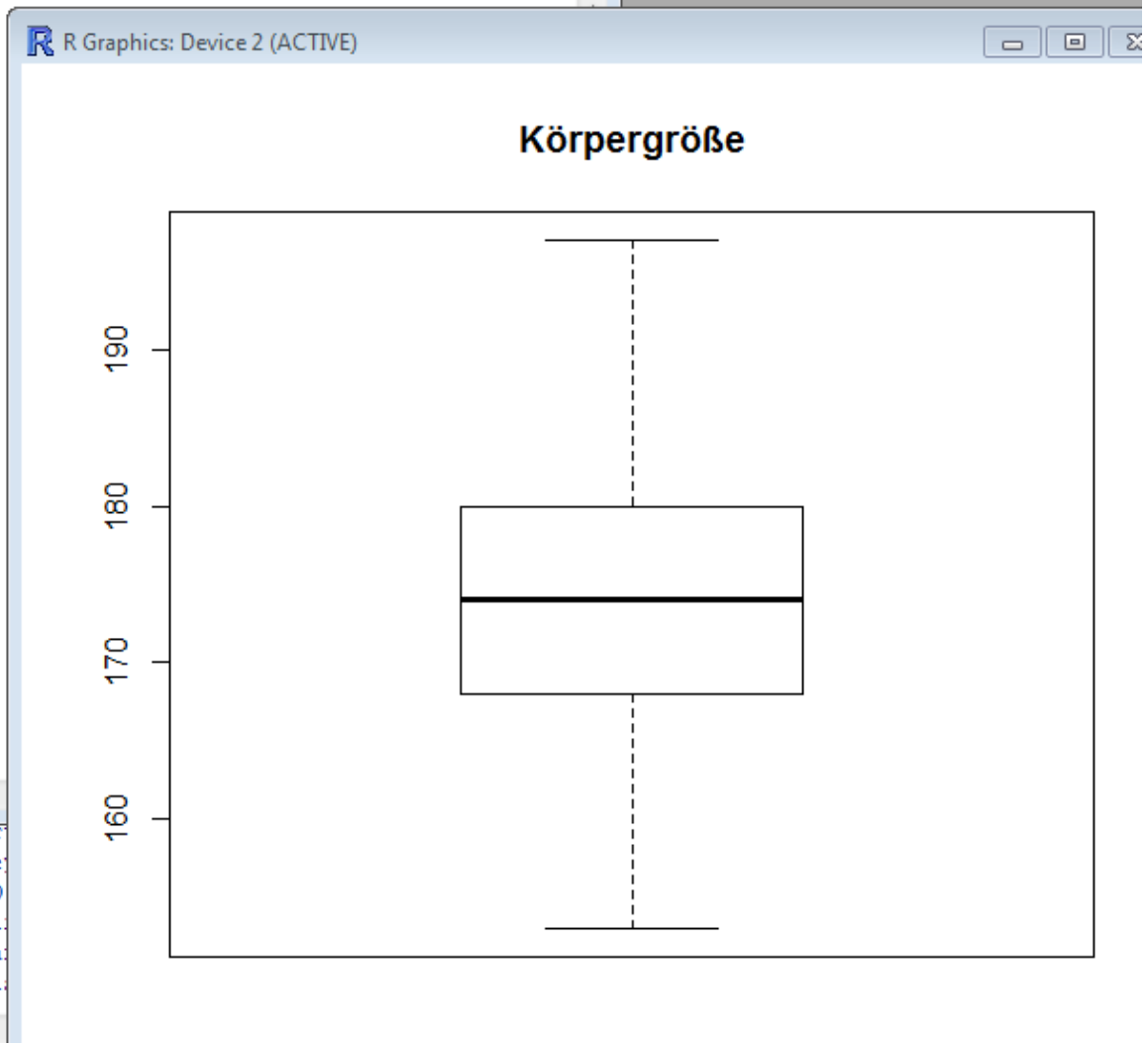




Boxplot für unseren Datensatz

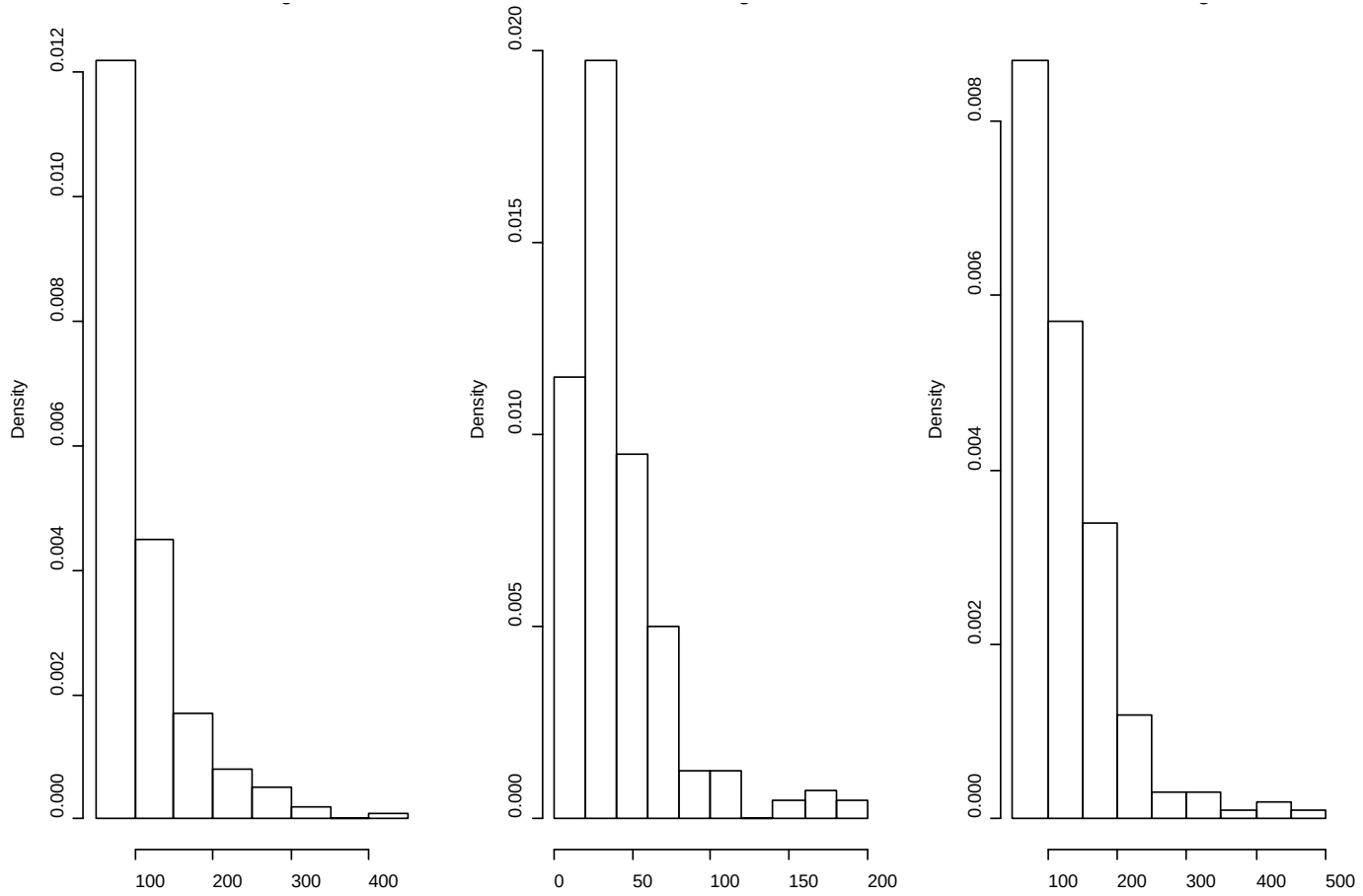
R Console

```
> # Boxplots  mit R
> boxplot(koerpergr, main="Körpergröße")
>
```



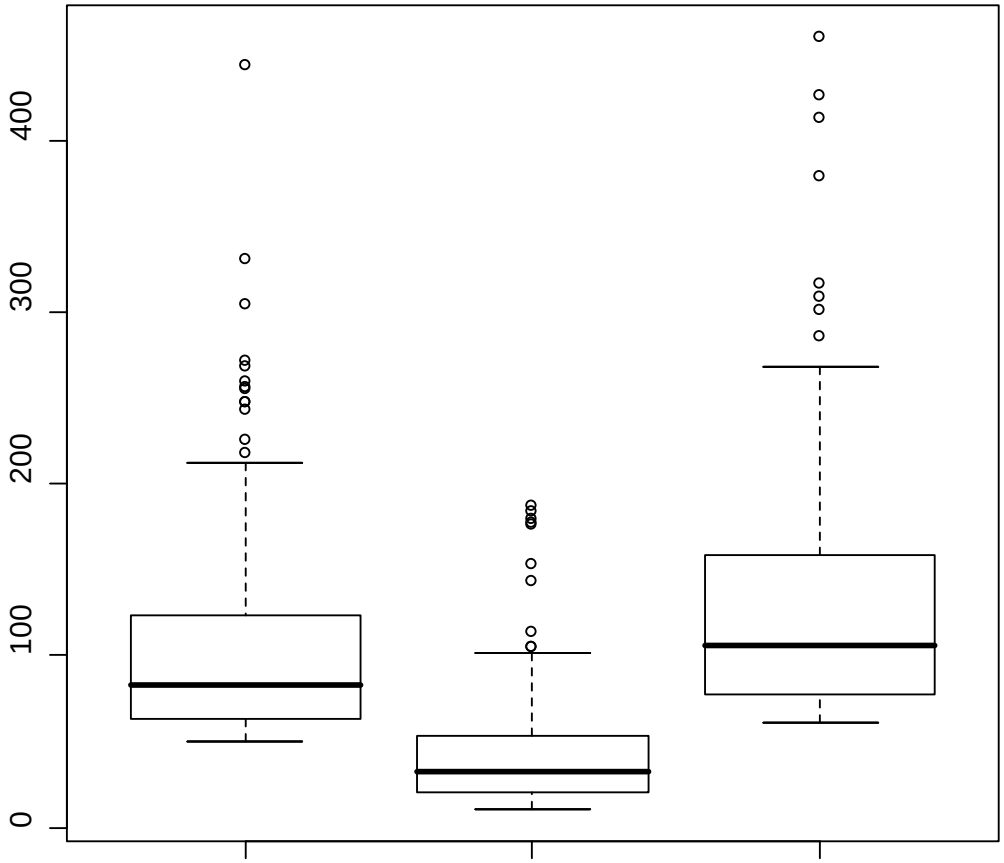
Vergleich von Verteilungen

Visualisierung mittels Histogrammen unterstützt die vergleichende Interpretation nur wenig



Vergleich von Verteilungen

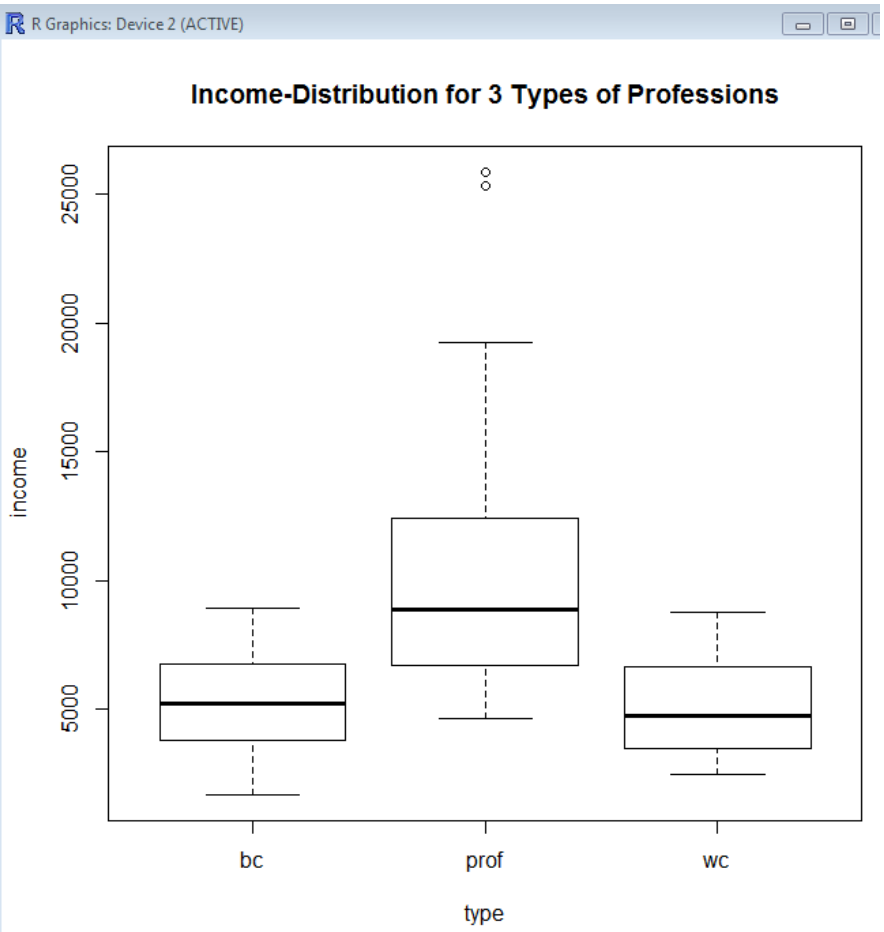
Visualisierung mittels Boxplots unterstützt die vergleichende Interpretation gut



Boxplots mit R in der Praxis

R Console

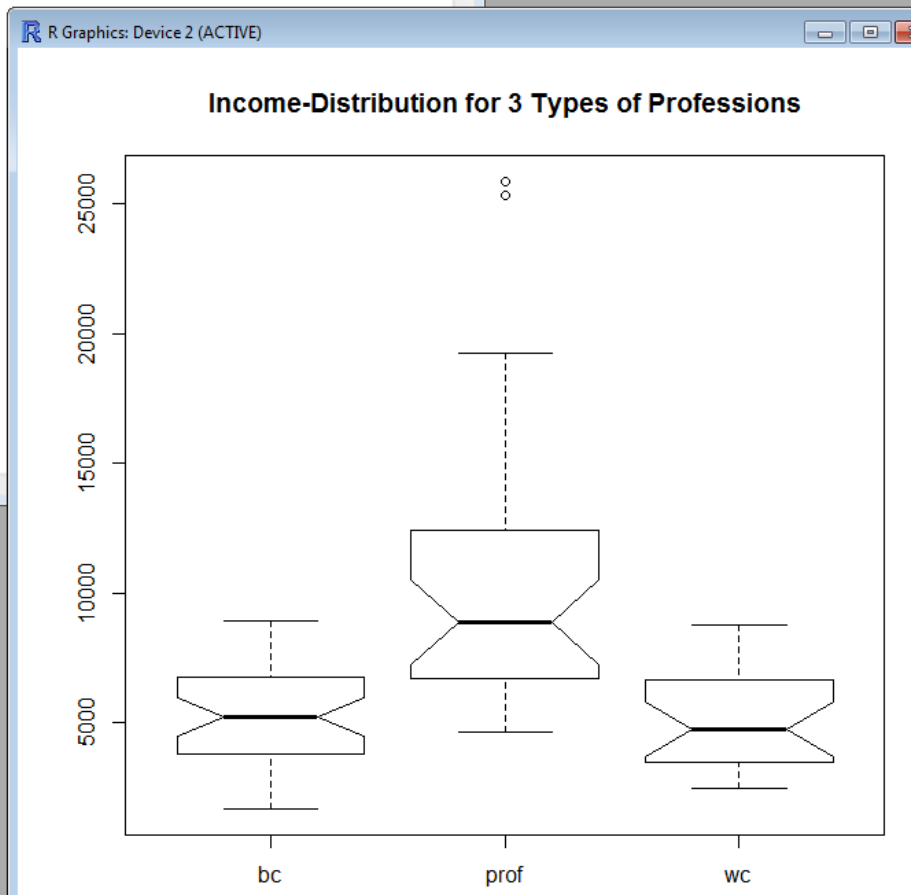
```
> # Beispiele zu Boxplots
> library(car)
> head(Prestige)
                  education income women prestige census type
gov.administrators     13.11  12351 11.16     68.8   1113 prof
general.managers       12.26  25879  4.02     69.1   1130 prof
accountants            12.77   9271 15.70     63.4   1171 prof
purchasing.officers    11.42   8865  9.11     56.8   1175 prof
chemists               14.62   8403 11.68     73.5   2111 prof
physicists             15.64  11030  5.13     77.6   2113 prof
> attach(Prestige)
Das folgende Objekt ist maskiert from package:datasets:

    women
> # Defaultkonzept der Darstellung stetiges versus diskretes Merkmal
> plot(income ~ type, main="Income-Distribution for 3 Types of Professions")
>
```



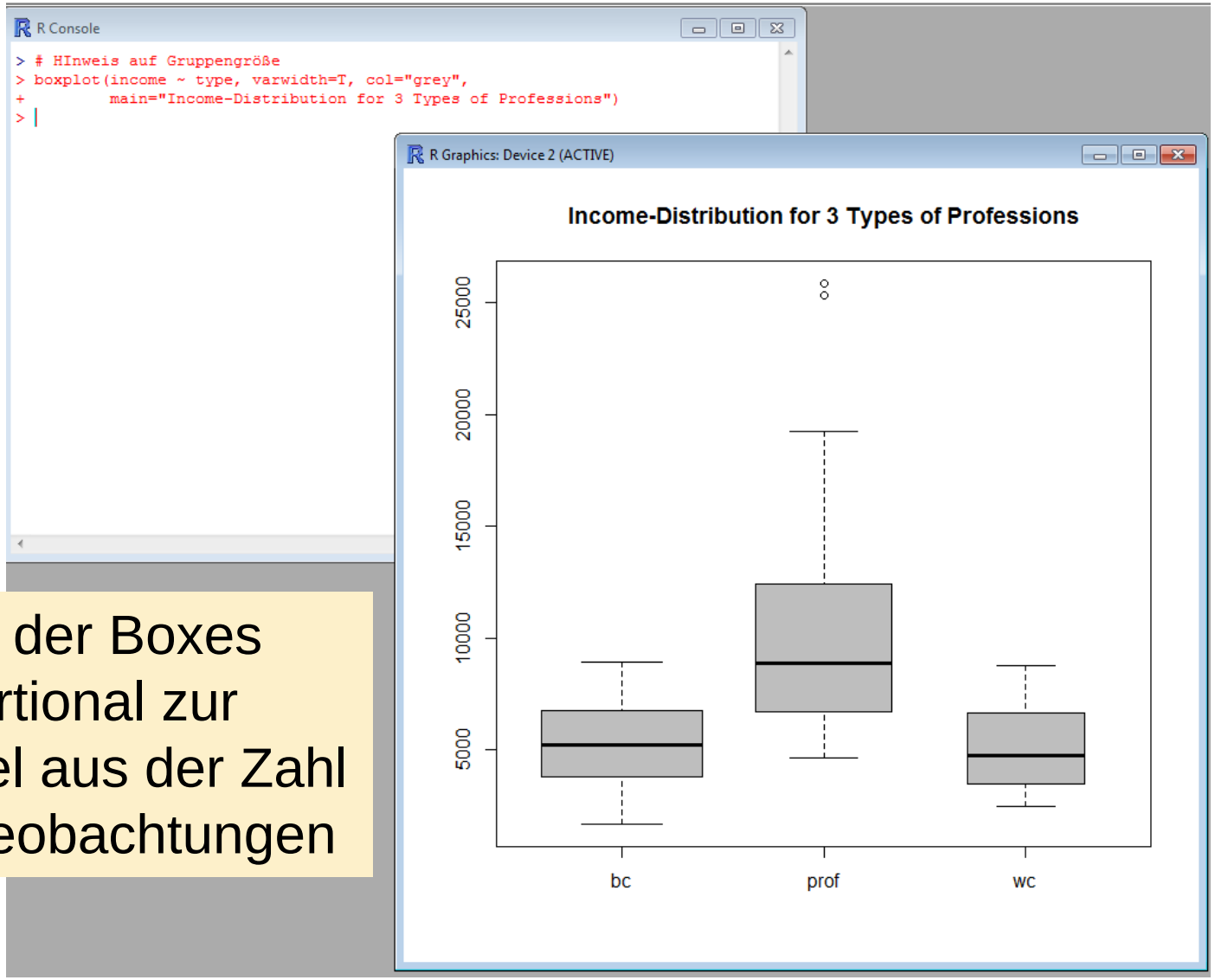
Notched Boxplot

R Console

```
> # Hinweis auf Unterschiedlichkeit der Mediane
> boxplot(income ~ type, notch=T,
+         main="Income-Distribution for 3 Types of Professions")
> |
```



Variable Width Boxplot



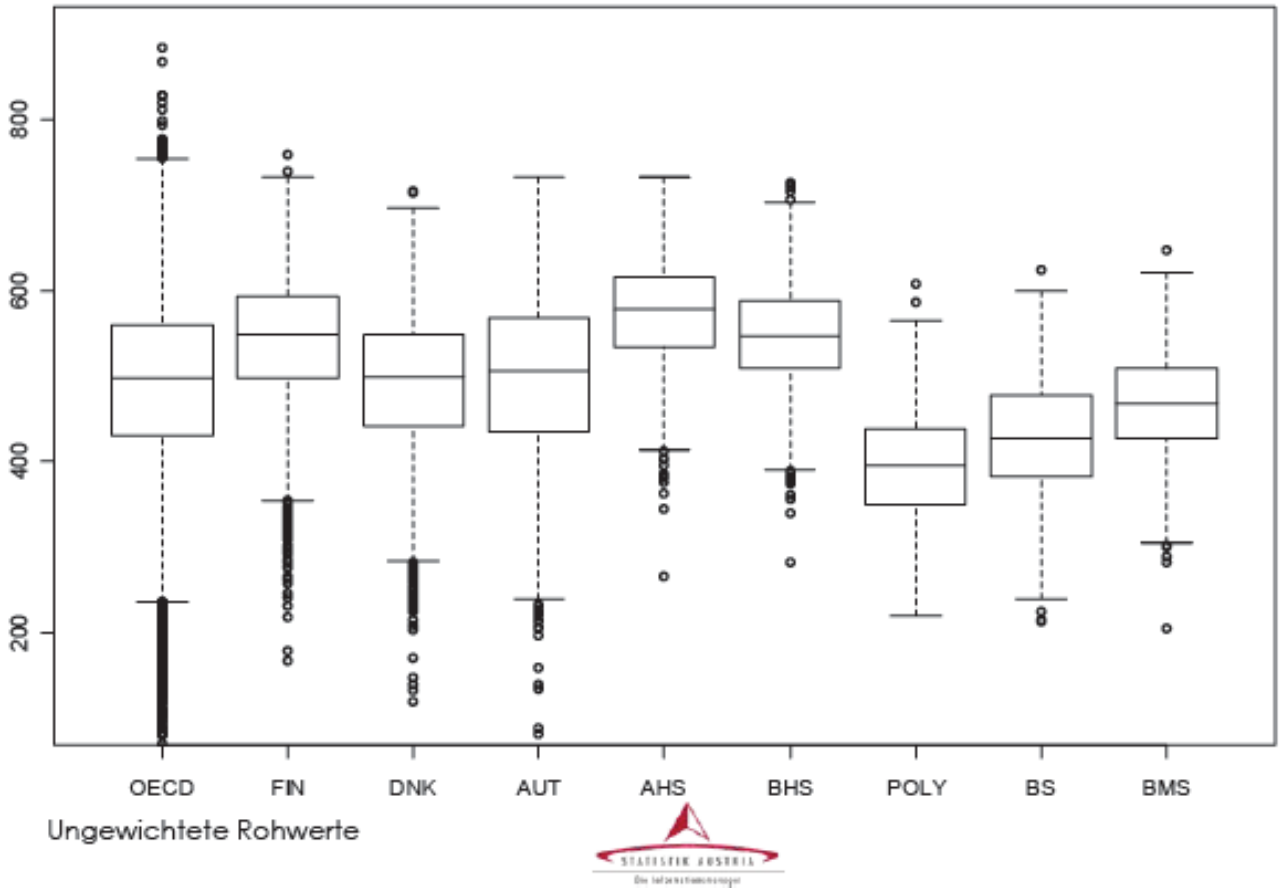
Breite der Boxes proportional zur Wurzel aus der Zahl der Beobachtungen

Vorteil von Boxplots

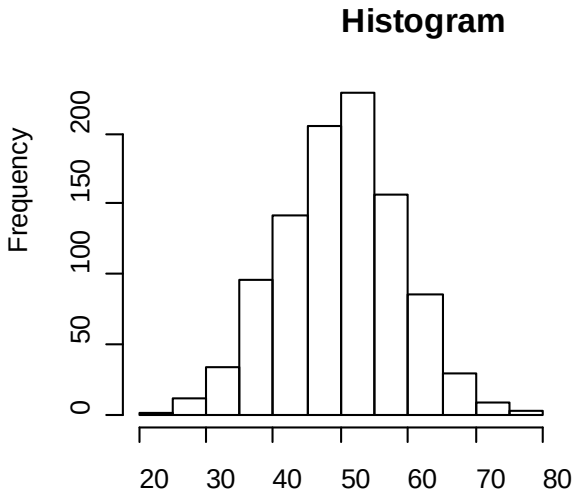
- Boxplots geben im Vergleich zum Histogramm zwar nur ein gröberes Bild von der Verteilung, aber sie sind viel besser zum Vergleich der Verteilung verschiedener Gruppen geeignet.

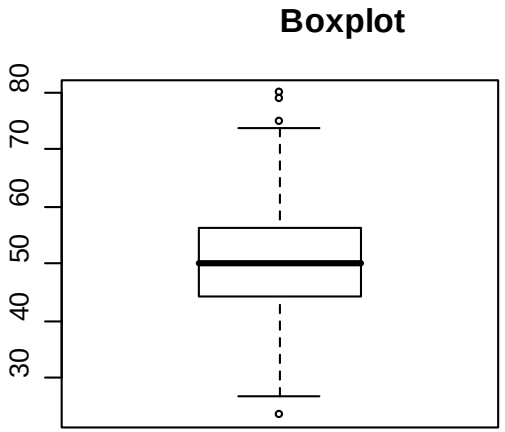
Pisa Studie 2000 – Darstellung von E. Neuwirth

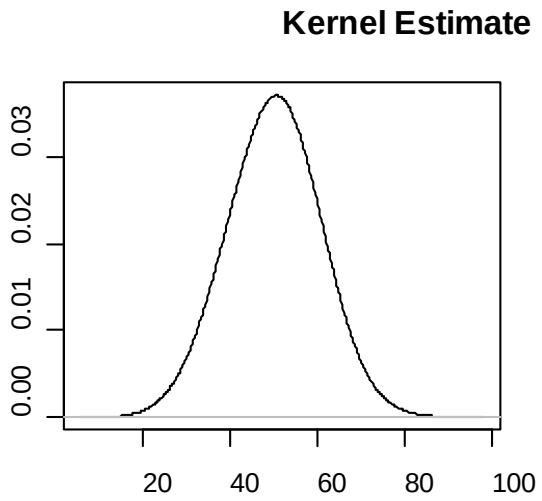
Lesefähigkeit

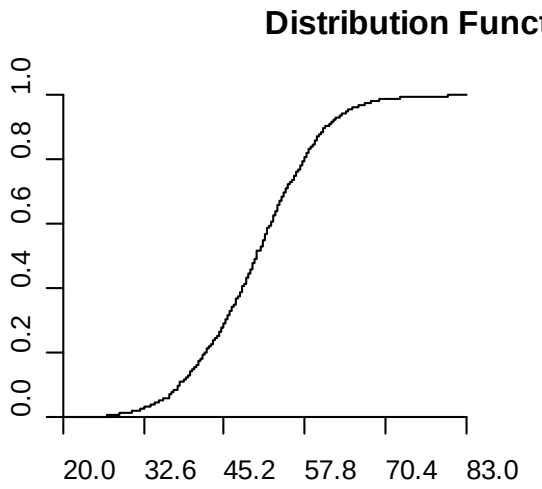


4 alternative Darstellungen eines Datensatzes

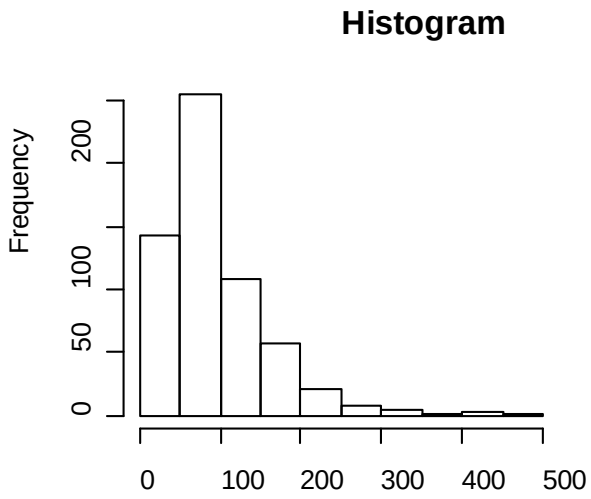


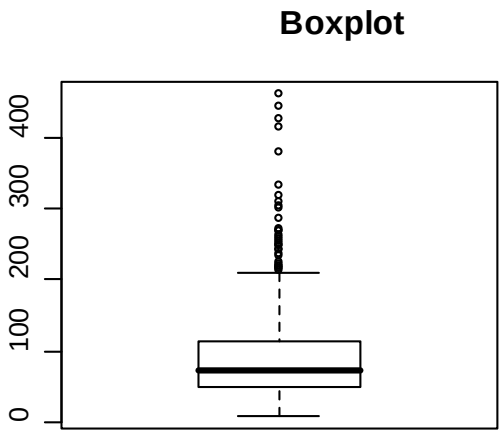


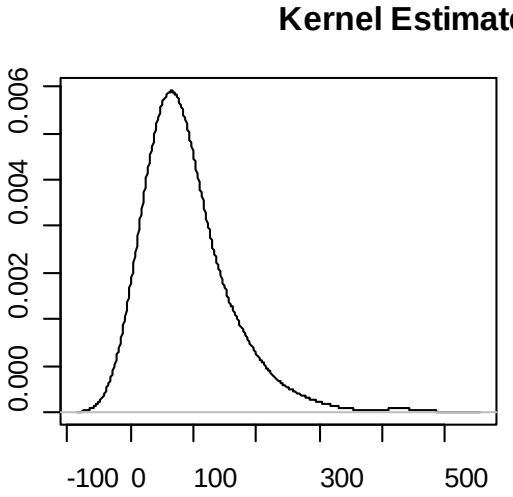


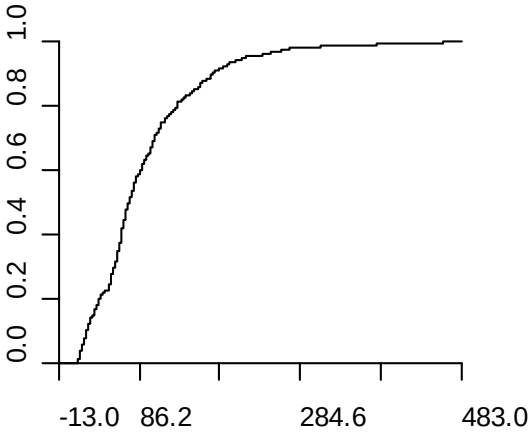


4 alternative Darstellungen eines Datensatzes

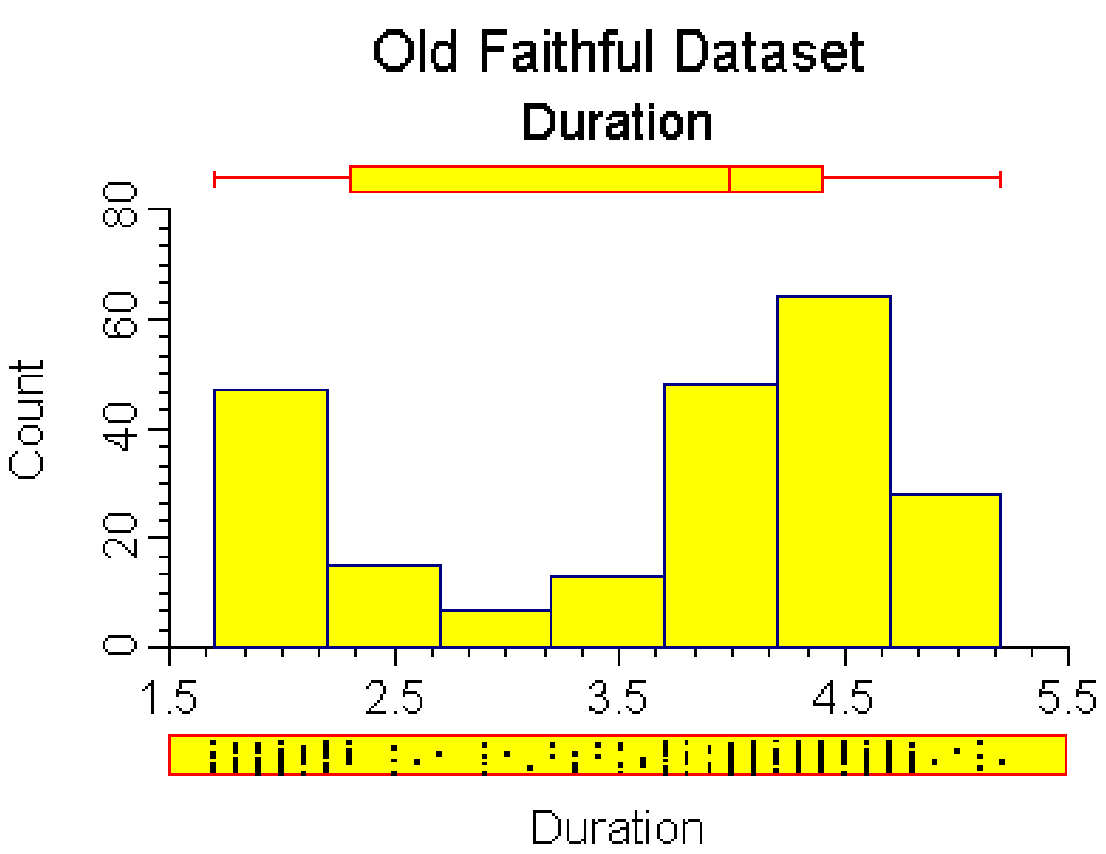








Kombination von Darstellungsvarianten



Die Kombination unterschiedlicher Darstellungstypen ist durchaus empfehlenswert, um die Form der Verteilung besser zu verstehen bzw. zu kommunizieren