4. Kumulierte Häufigkeiten und Quantile

Kumulierte Häufigkeiten

- Oft ist man nicht an der Häufigkeit einzelner Merkmalsausprägungen interessiert, sondern an der Häufigkeit von Intervallen.
- Typische Fragestellung:
  - Wie groß ist der Anteil aller Merkmalsträger mit einem Merkmalswert größer (bzw. kleiner) als ein bestimmter Wert x?
- Hierzu summiert man die Häufigkeitstabelle schrittweise auf.
- Hinweis:
  Sinnvolle Kumulation erfordert, dass das Merkmal zumindest ordinal skaliert ist!

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		

Kumulierte Häufigkeiten (klassierte Daten)

Bereich	ni	hi	Ni	Hi
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		

Kumulierte Häufigkeiten

- Die absoluten kumulierten relativen Häufigkeiten geben an, wie viele Beobachtungen einen bestimmten Wert x nicht übertreffen.

  $N(X \leq x)$ z.B. 56 Studenten sind kleiner gleich 175 cm

- Die entsprechenden relativen kumulierten Häufigkeiten bezeichnen wir mit

  $H(X \leq x)=N(X \leq x)/n$ z.B. 76% der Studenten sind kleiner gleich 180 cm

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

- Die empirische Verteilungsfunktion F(x) ist definiert durch

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

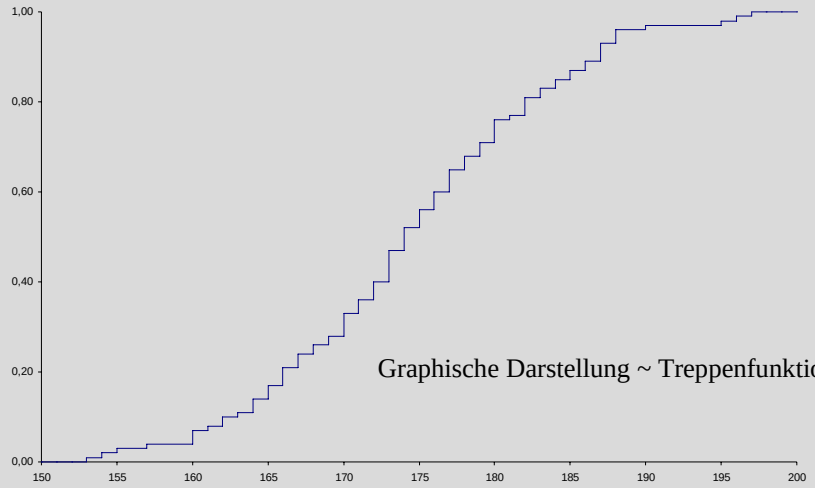
Kumulierte Häufigkeiten (Einzeldaten)

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

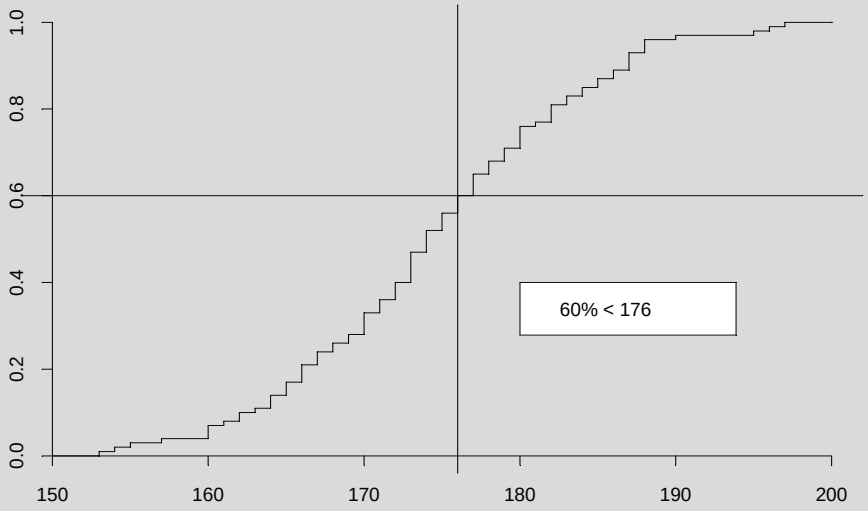
Kumulierte relative Häufigkeiten ~ Empirische Verteilungsfunktion

Empirische Verteilungsfunktion



Graphische Darstellung ~ Treppenfunktion

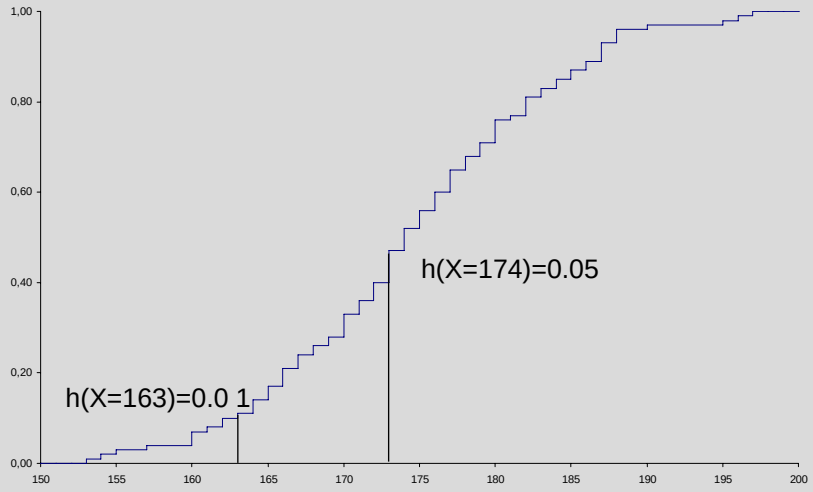
Empirische Verteilungsfunktion (Leseprobe)



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

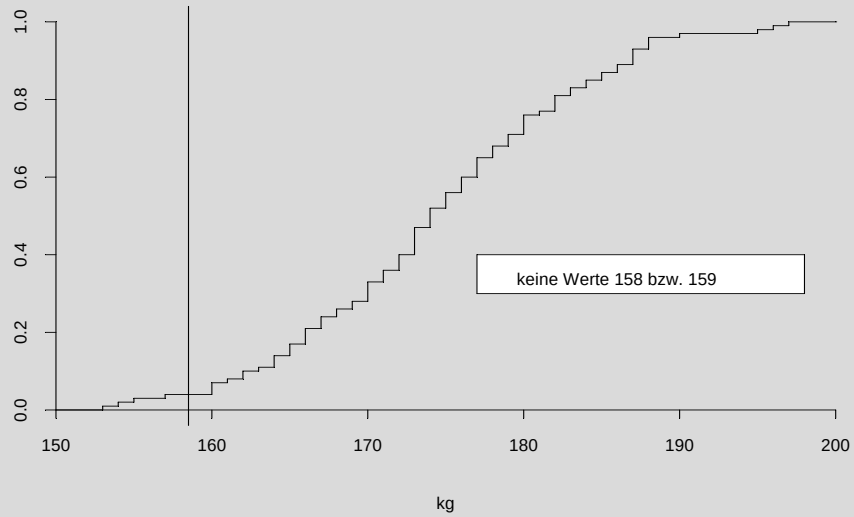
Unterschiedliche Sprunghöhen



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 empirische Verteilungsfunktion in diesem Bereich horizontal

Empirische Verteilungsfunktion



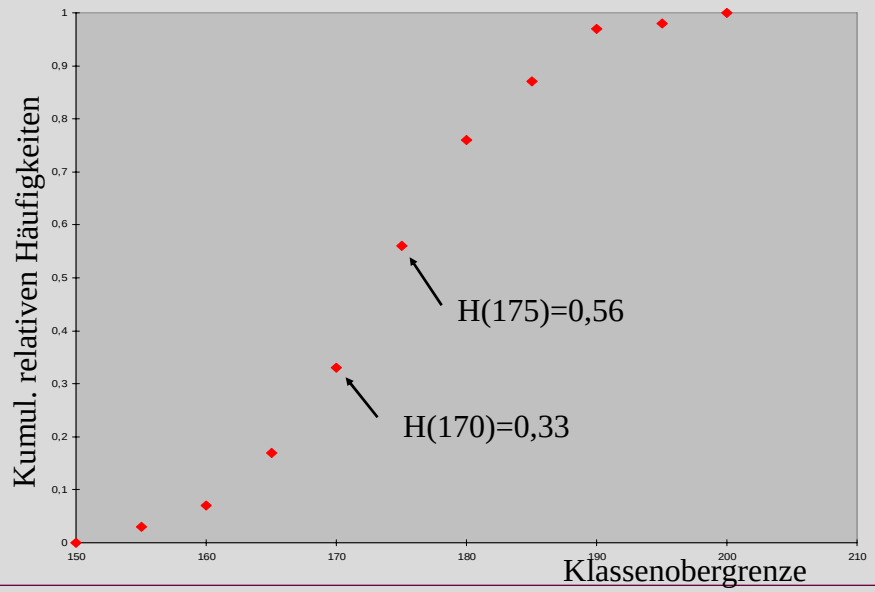
Eigenschaften der emp. Verteilungsfunktion

- 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		

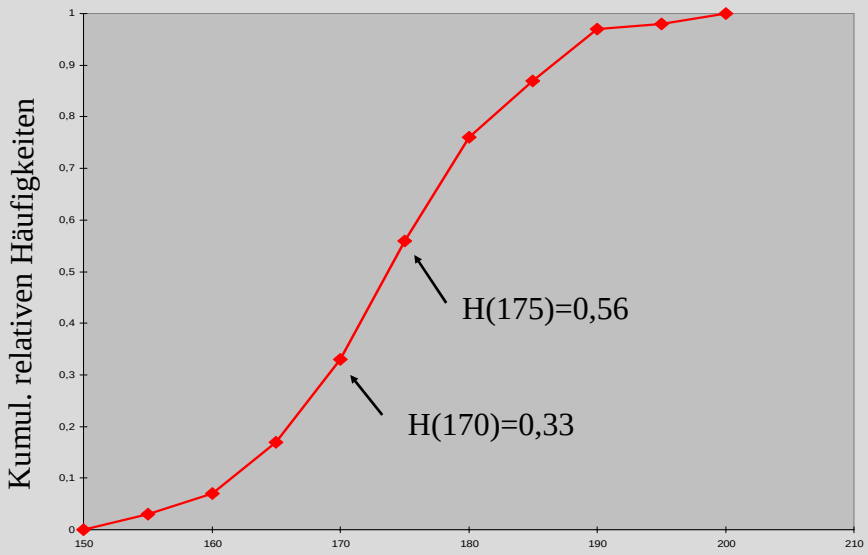
Verteilungsfunktion bei klassierten Daten



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 in der Klasse

Summenkurve

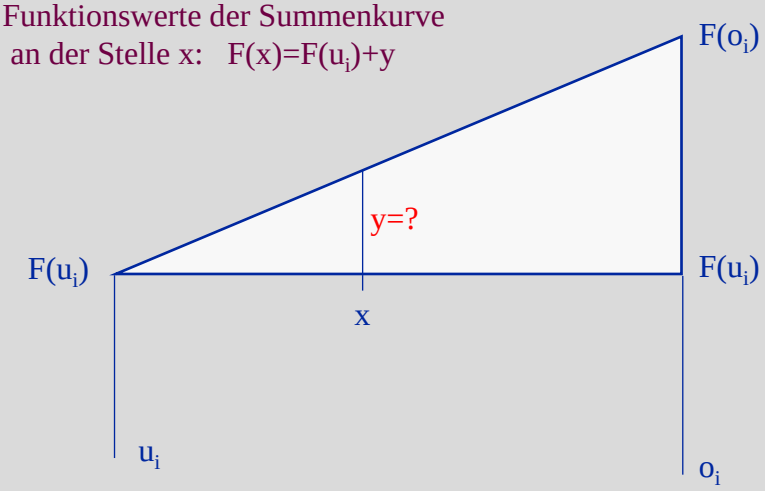


Verteilungsfunktion bei klassierten Daten (Beispiel)

- 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

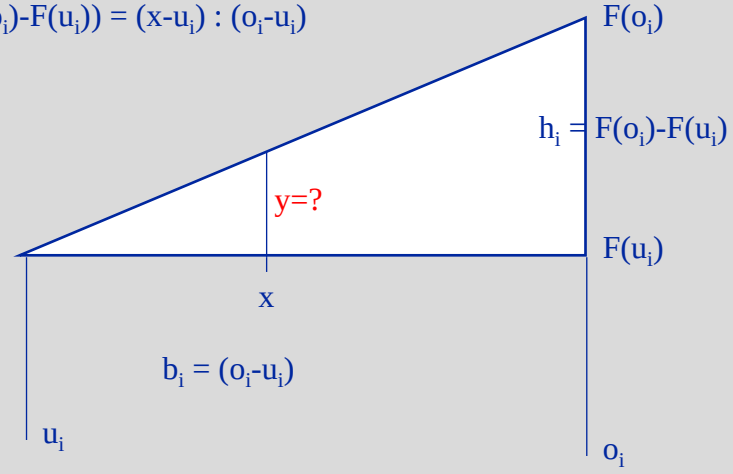
Interpolation

Funktionswerte der Summenkurve
an der Stelle x: $F(x)=F(u_i)+y$



Anwendung des Strahlensatzes

$y : (F(o_i)-F(u_i)) = (x-u_i) : (o_i-u_i)$

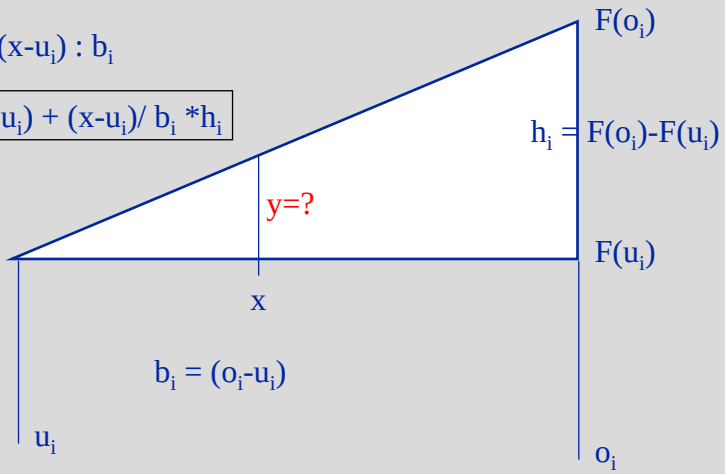


Anwendung des Strahlensatzes

$y : (F(o_i)-F(u_i)) = (x-u_i) : (o_i-u_i)$

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

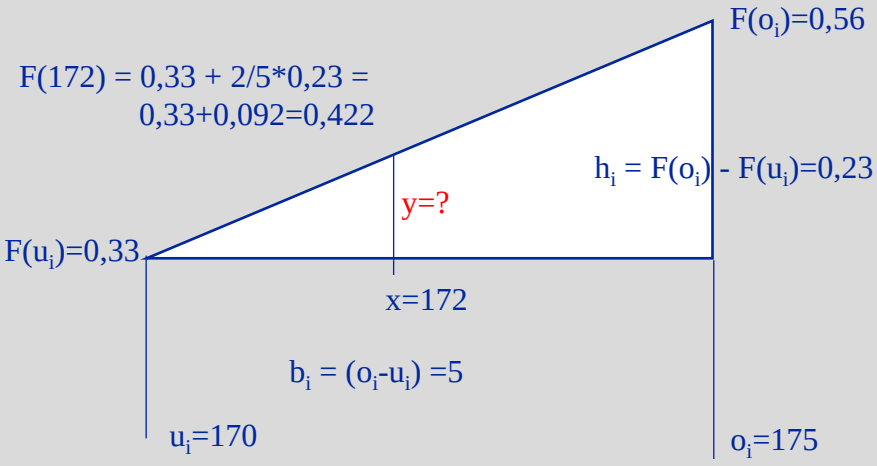
$F(x) = F(u_i) + (x-u_i)/ b_i * h_i$



Beispiel

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

$F(172) = 0,33 + 2/5*0,23 =$
$0,33+0,092=0,422$

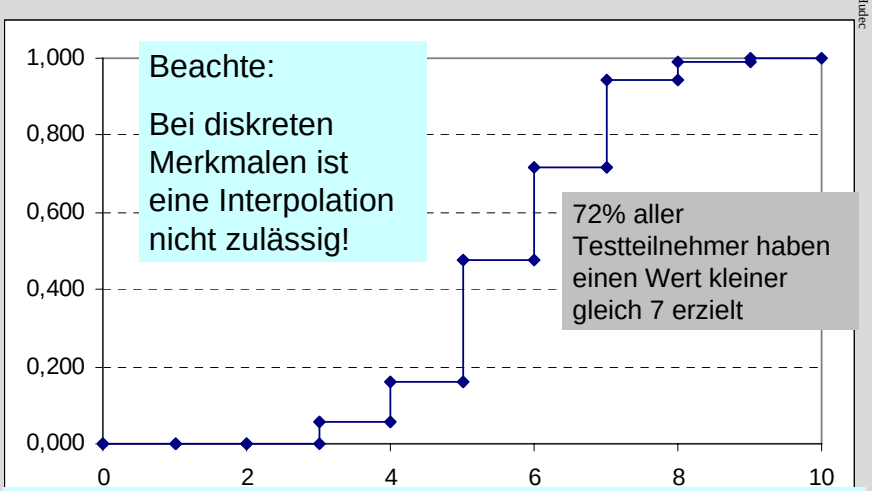


Empirische Verteilungsfunktion bei diskreten Merkmalen

Beispiel: "Produktives Denken"

i	x_i	h_i	$H(X \leq x_i)$
1	0	0,00	0,00
2	1	0,00	0,00
3	2	0,00	0,00
4	3	0,06	0,06
5	4	0,10	0,16
6	5	0,32	0,48
7	6	0,24	0,72
8	7	0,23	0,94
9	8	0,05	0,99
10	9	0,01	1,00
Gesamt		1	

Empirische Verteilungsfunktion "Produktives Denken"



Hier gibt nur die Treppenkurve ein korrektes Bild der Verteilung

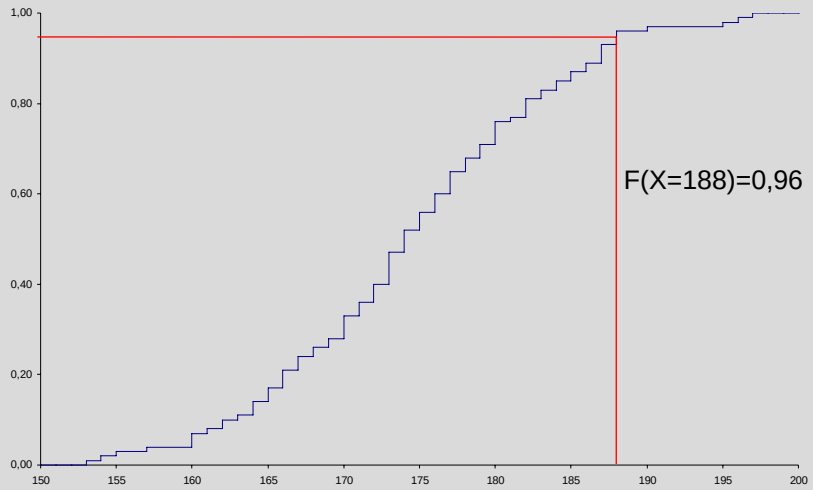
Empirische Quantile

- Ausgehend von einem Anteilswert p (y-Achse) wird der zugehörige Wert bestimmt, für den F(x) zum erstenmal größer als oder zumindest gleich groß wie p ist.
- Das bedeutet ein p-Quantil ist jener möglichst kleine Merkmalswert für den gerade noch gilt, dass p-Prozent der Beobachtungen kleiner gleich als eben dieser Merkmalswert sind.
- $0 < p < 1$
- Datensatz: $x_1, ..., x_n$
- Das Empirische p-Quantil x_p ist der kleinste Wert x für den $F(x) \geq p$ gilt.
- Seien $x_{(1)}, ... x_{(n)}$ die geordneten Werte:
- $x_p = x_{(k)}$, wobei k wie folgt gegeben ist:
  $(k-1)/n < p \leq k/n$

Beispiele zu empirischen Quantilen

- Gesucht ist ein Wert, so dass 95% der Studenten kleiner gleich diesem Wert sind
- Datensatz Körpergröße n=100
  $x_{0,95}$ = ?
  $(k-1)/n < p \leq k/n \rightarrow (k-1) < np \leq k$
  $(k-1) < 95 \leq k \Longrightarrow k=95$
  $x_{0,95} = 188$
- Datensatz produktives Denken n=120
  $x_{0,50}$ = ?
  $(k-1) < 60 \leq k \Longrightarrow k=60$
  $x_{0,50} = 7$

Bestimmung des Quantils



Wichtige Quantile

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 $x_{0,1}$ ... $x_{0,9}$

Empirische Quantile

Beispiel: Körpergröße (Originalwerte)

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

Five Numbers Summary

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

Empirische Quantile bei klassierten Daten

- Bei klassierten Daten ergibt sich das p-Quantil durch Interpolation
- Ausgangspunkt ist jene Klasse, in der die kumulierten Häufigkeiten den p-Wert übersteigen

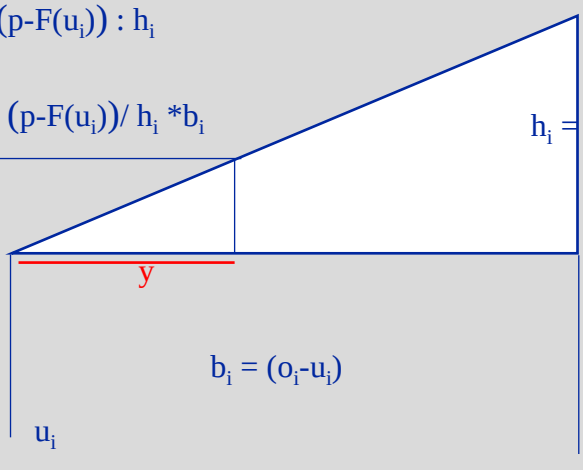
Bestimmung des 0,5 Quantils

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		

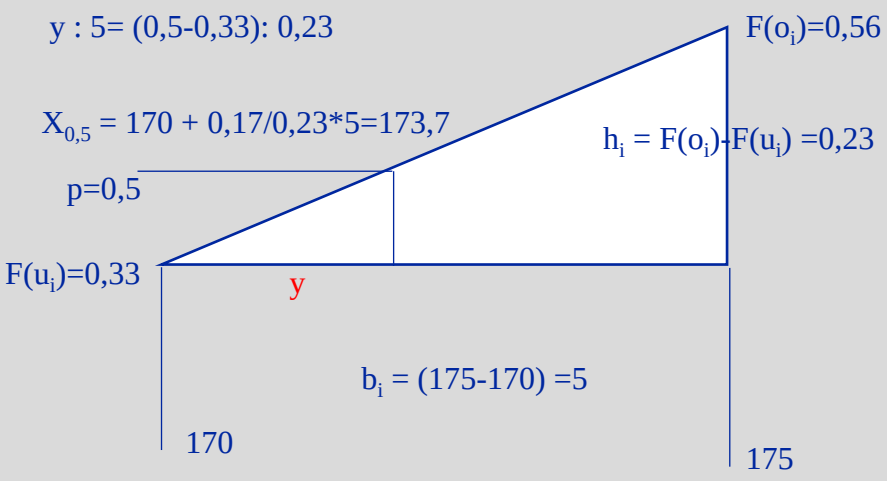
Empirische Quantile bei klassierten Daten

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

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



Empirische Quantile bei klassierten Daten

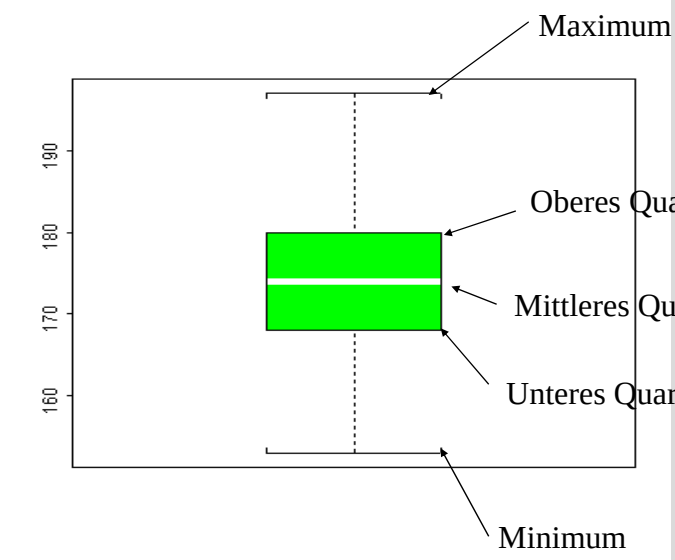


Box-Plots

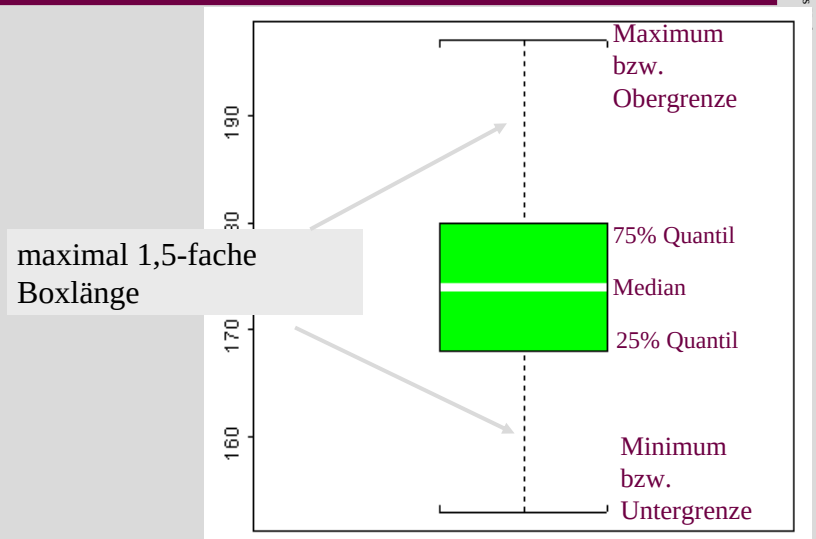
- Basierend auf den 5 zusammenfassenden Werten einer Verteilung:
  - Minimum, 1.Quartil, 2.Quartil, 3.Quartil und Maximum

  lassen sich instruktive Graphiken zur Darstellung einer Verteilung entwickeln, die insbesondere zum Vergleich mehrerer Gruppen gut geeignet sind.
- Häufig werden die horizontal 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.

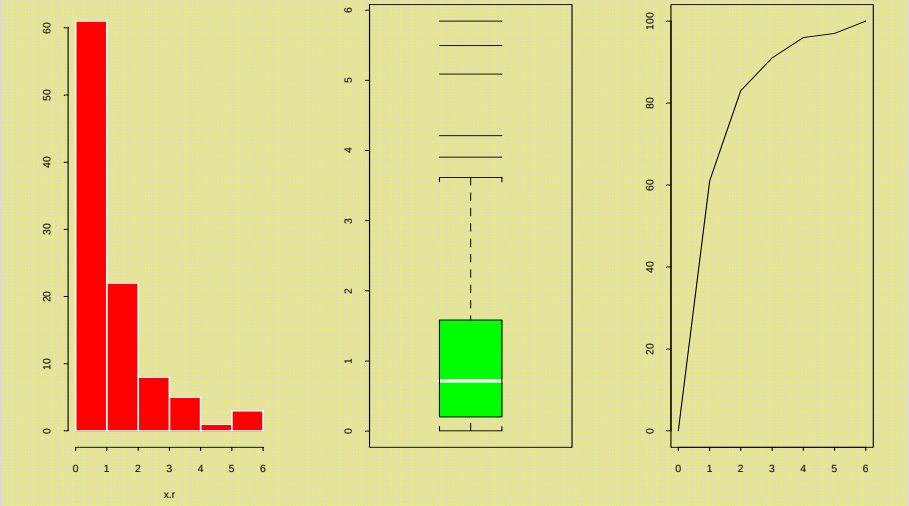
Boxplot



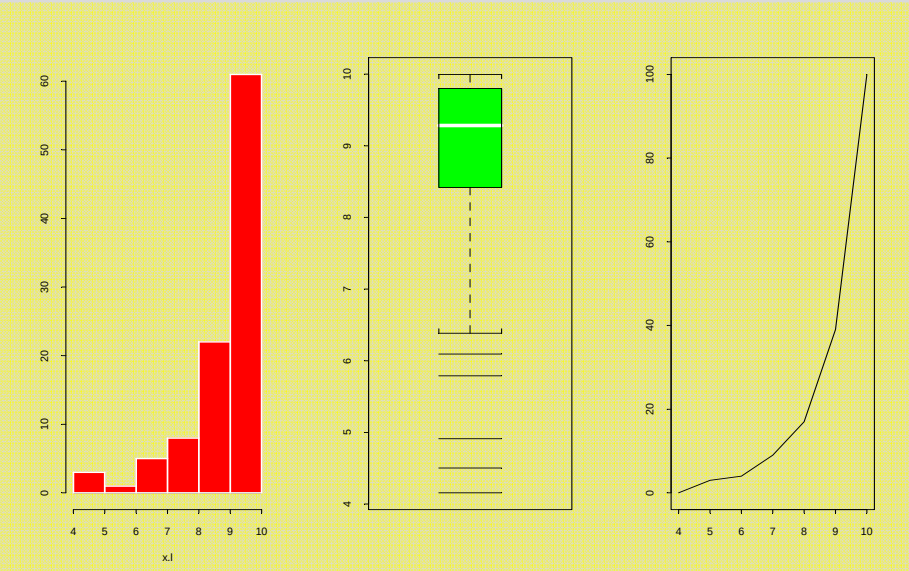
Boxplot



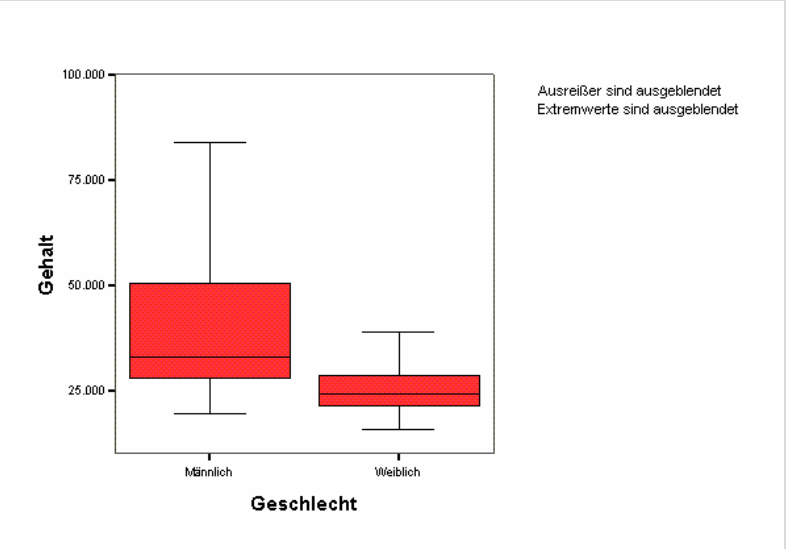
Beispiel einer rechtsschiefen Verteilung



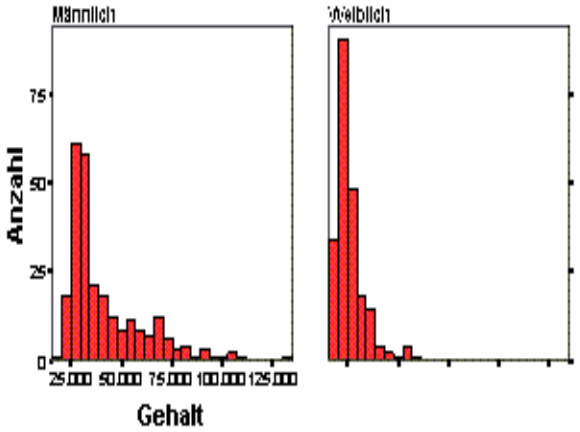
Beispiel einer linksschiefen Verteilung



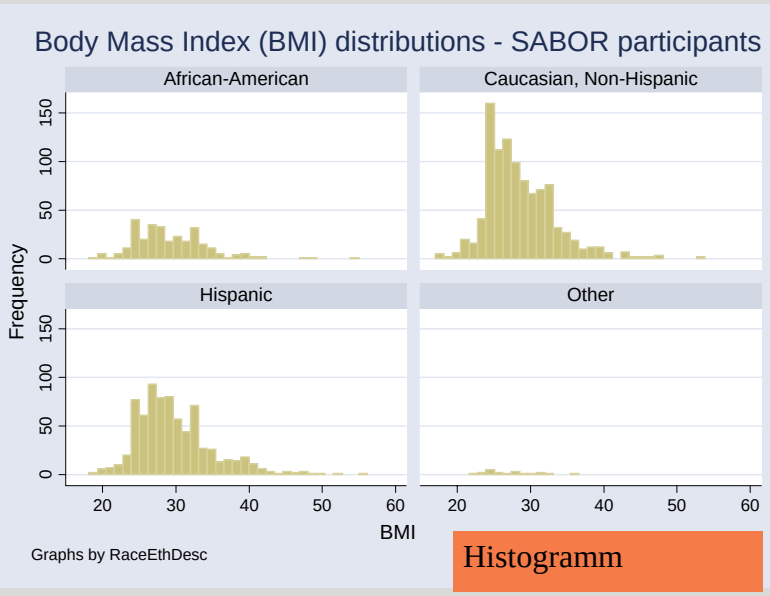
Vergleich zweier Verteilungen



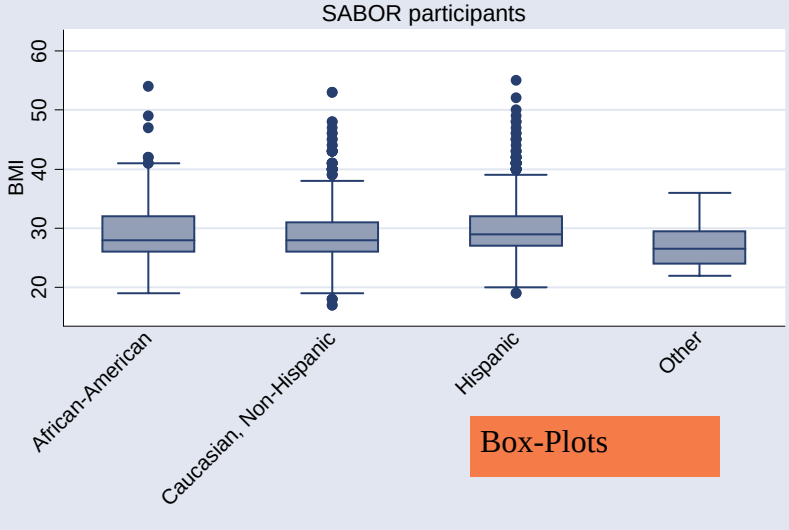
Vergleich zweier Verteilungen



Insbesondere beim Vergleich von Verteilungen ist der Boxplot oft übersichtlicher als das Histogramm



Histogramm



Box-Plots