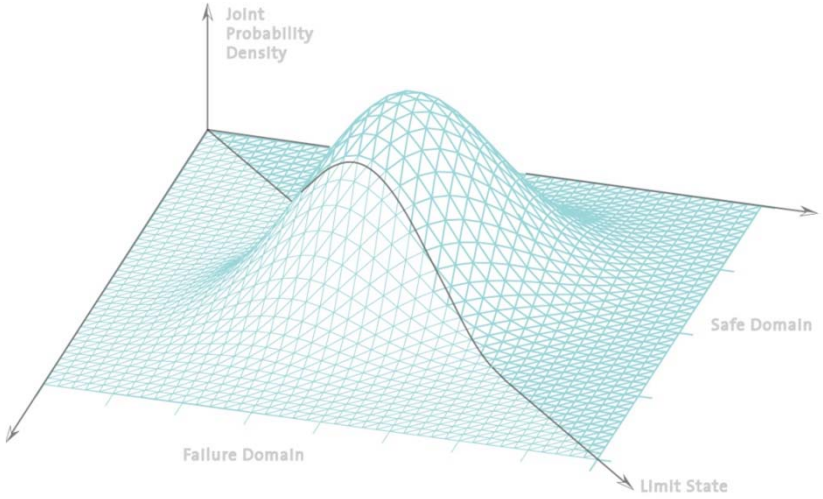


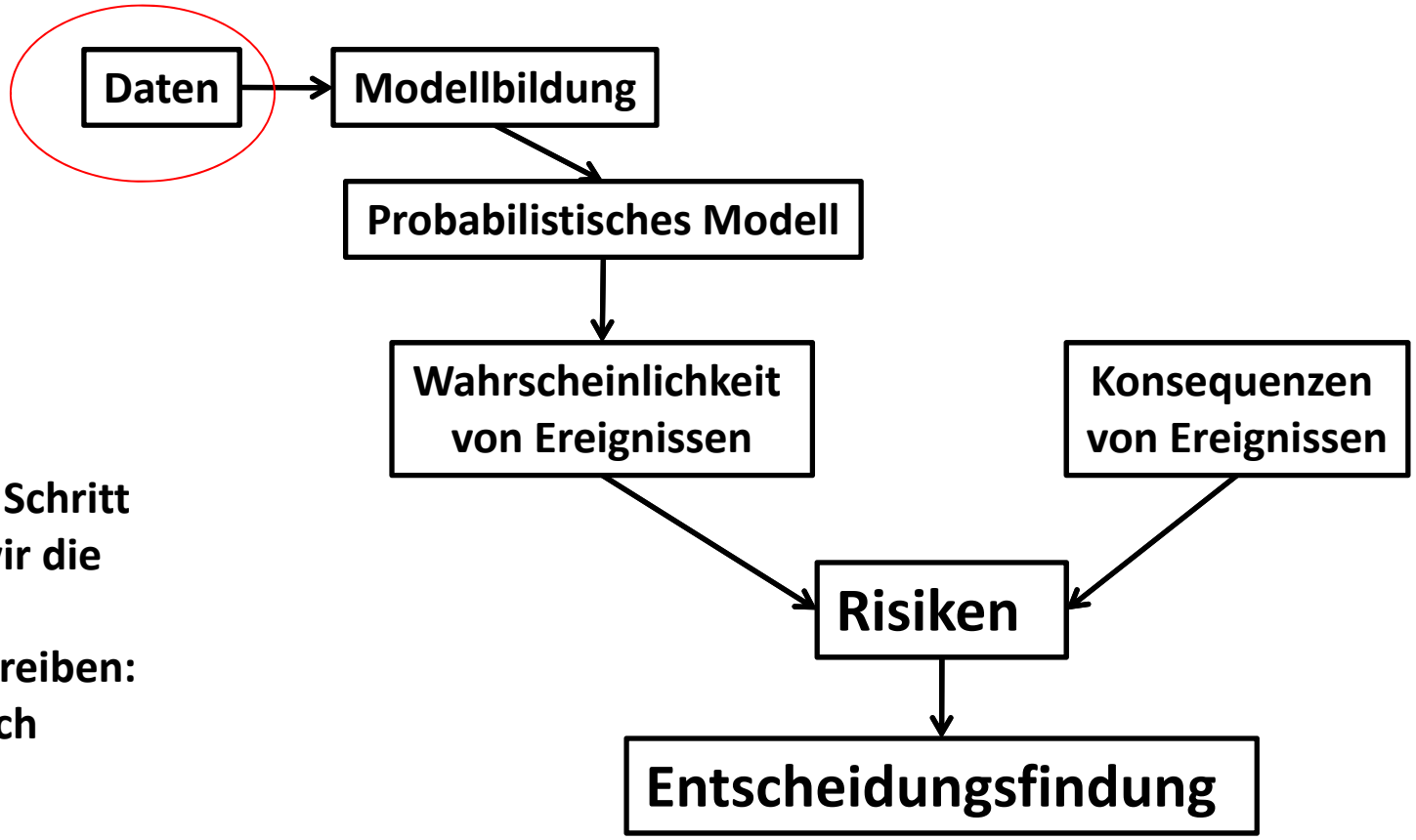
Statistik und Wahrscheinlichkeitsrechnung

3. Vorlesung

Dr. Jochen Köhler

Inhalte der heutigen Vorlesung

- Ziel:



Im ersten Schritt werden wir die Daten nur beschreiben:
- numerisch
- grafisch

Inhalte der heutigen Vorlesung

- Überblick der beschreibenden Statistik
- **Numerische Kennwerte**
  Mit welchen einfachen Zahlen können Datenmengen charakterisiert werden?
- **Grafische Darstellung von Datenmengen**
  Wie werden Datenmengen informativ in Grafiken umgesetzt?

Ziel der beschreibenden Statistik

- Beschreiben von Datenmengen

Körpergrösse

186	183	180	183	178	183	183	189	189	168	172	165
173	162	182	178	185	185	179	187	186	178	165	170
185	178	172	186	171	190	179	175	160	160	157	177
183	173	180	180	183	187	170	182	179	179	172	169
170	184	180	179	175	183	165	169	193	176	154	
180	169	179	188	180	170	183	180	176	165	170	
175	183	183	173	183	175	178	176	188	168	156	
180	182	186	184	178	190	160	179	176	166	175	
180	177	200	172	177	179	190	182	195	167	162	
182	173	175	175	180	182	174	186	170	172	159	
192	184	183	176	183	185	175	186	170	164	170	
188	177	174	182	187	170	185	178	172	168	168	
177	186	178	178	175	182	188	187	164	178	169	
174	180	180	188	182	188	192	193	171	166	162	
187	180	184	178	186	183	173	180	173	171	175	
176	187	185	174	183	190	175	172	168	164	170	
176	188	178	176	177	180	184	188	177	173	168	
180	192	181	177	182	188	178	180	169	180	170	
175	179	179	188	170	185	180	182	155	174	166	
176	183	172	180	184	188	186	185	167	170	169	

⇨ **Kennwerte**

⇨ **Grafiken**

Keine Annahmen –
nur Beschreibung !!

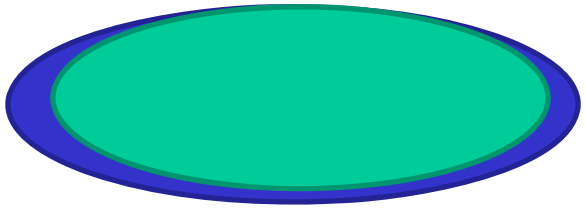
Vorbemerkung

- Stichprobe und Grundgesamtheit

  – Die statistischen Eigenschaften einer Grundgesamtheit werden anhand von Stichproben untersucht.

  Z.B.: Die Grundgesamtheit aller Studierenden, welche für Statistik und Wahrscheinlichkeitsrechnung eingeschrieben sind, ist m = 308.

  Stichprobe von letzter Woche, n = 224.



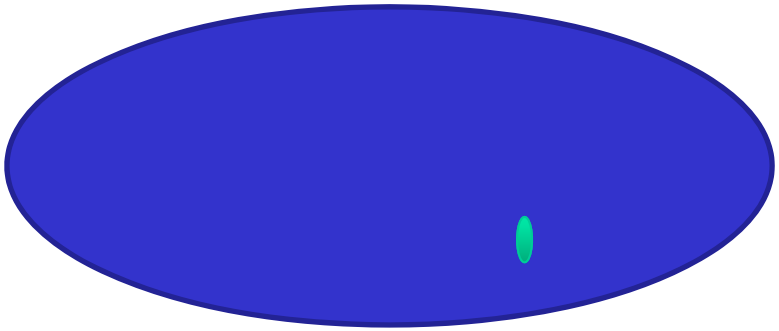
Vorbemerkung

- Stichprobe und Grundgesamtheit

  – Die statistischen Eigenschaften einer Grundgesamtheit werden anhand von Stichproben untersucht.

  Z.B.: Biegezähigkeit von Büroklammern, $m = \infty$.

  Stichprobe, $n = 222$



Vorbemerkung

- Stichprobe und Grundgesamtheit
  - Die statistischen Eigenschaften einer Grundgesamtheit werden anhand von Stichproben untersucht.
  - Damit die Stichprobe die Grundgesamtheit repräsentiert, müssen die Stichproben **zufällig** aus der Grundgesamtheit entnommen werden.

Ziel der beschreibenden Statistik

- Beschreiben von Datenmengen

Körpergrösse

186	183	180	183	178	183	183	189	189	168	172	165
173	162	182	178	185	185	179	187	186	178	165	170
185	178	172	186	171	190	179	175	160	160	157	177
183	173	180	180	183	187	170	182	179	179	172	169
170	184	180	179	175	183	165	169	193	176	154	
180	169	179	188	180	170	183	180	176	165	170	
175	183	183	173	183	175	178	176	188	168	156	
180	182	186	184	178	190	160	179	176	166	175	
180	177	200	172	177	179	190	182	195	167	162	
182	173	175	175	180	182	174	186	170	172	159	
192	184	183	176	183	185	175	186	170	164	170	
188	177	174	182	187	170	185	178	172	168	168	
177	186	178	178	175	182	188	187	164	178	169	
174	180	180	188	182	188	192	193	171	166	162	
187	180	184	178	186	183	173	180	173	171	175	
176	187	185	174	183	190	175	172	168	164	170	
176	188	178	176	177	180	184	188	177	173	168	
180	192	181	177	182	188	178	180	169	180	170	
175	179	179	188	170	185	180	182	155	174	166	
176	183	172	180	184	188	186	185	167	170	169	

⇨ **Kennwerte**

⇨ **Grafiken**

Keine Annahmen –
nur Beschreibung !!

Datenbeschreibung

- Zusammenfassen zu nur einem Kennwert

Arithmetisches Mittel: $$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

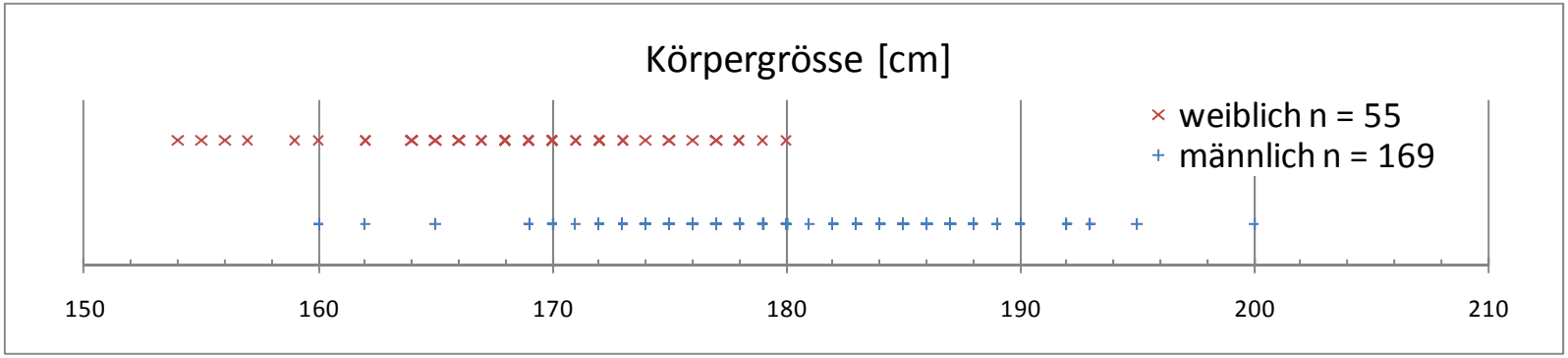
Für einen Datensatz: $$\mathbf{x} = \left(x_1, x_2, \ldots, x_n\right)^T$$

Um eine Stichprobe nur mit Hilfe eines Kennwertes zu beschreiben, wird normalerweise der Stichproben-Mittelwert verwendet.

Datenbeschreibung

- Einfache grafische Darstellung von Stichproben

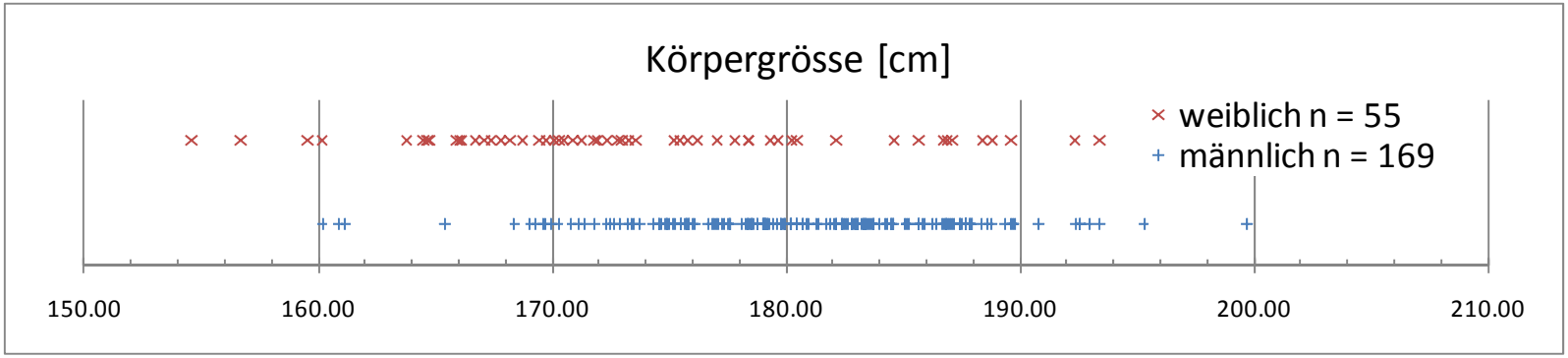
Eindimensionales Streudiagramm:



Datenbeschreibung

- Einfache grafische Darstellung von Stichproben

Eindimensionales Streudiagramm:



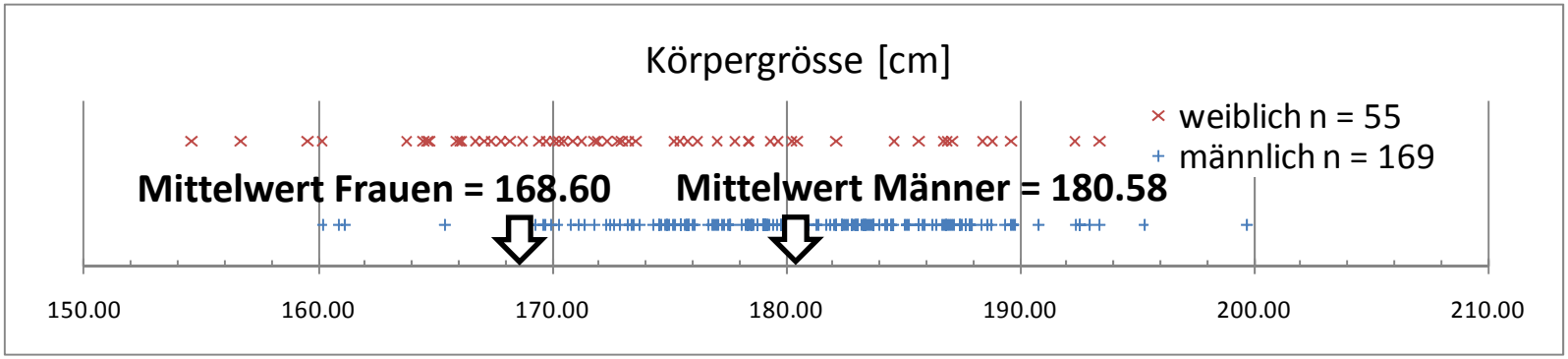
Guter Datenüberblick (Maximum, Minimum).

Vorsicht bei diskret verteilten Daten !

Datenbeschreibung

- Einfache grafische Darstellung von Stichproben

Eindimensionales Streudiagramm:



Der Stichprobenmittelwert $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ entspricht dem „Schwerpunkt“ der Daten.

Datenbeschreibung

- Einfache grafische Darstellung von Stichproben

Histogramm:

Einteilung der Datenreihe in Intervalle.

Darstellung der Grösse der Intervalle.

z.B. die Körpergrösse

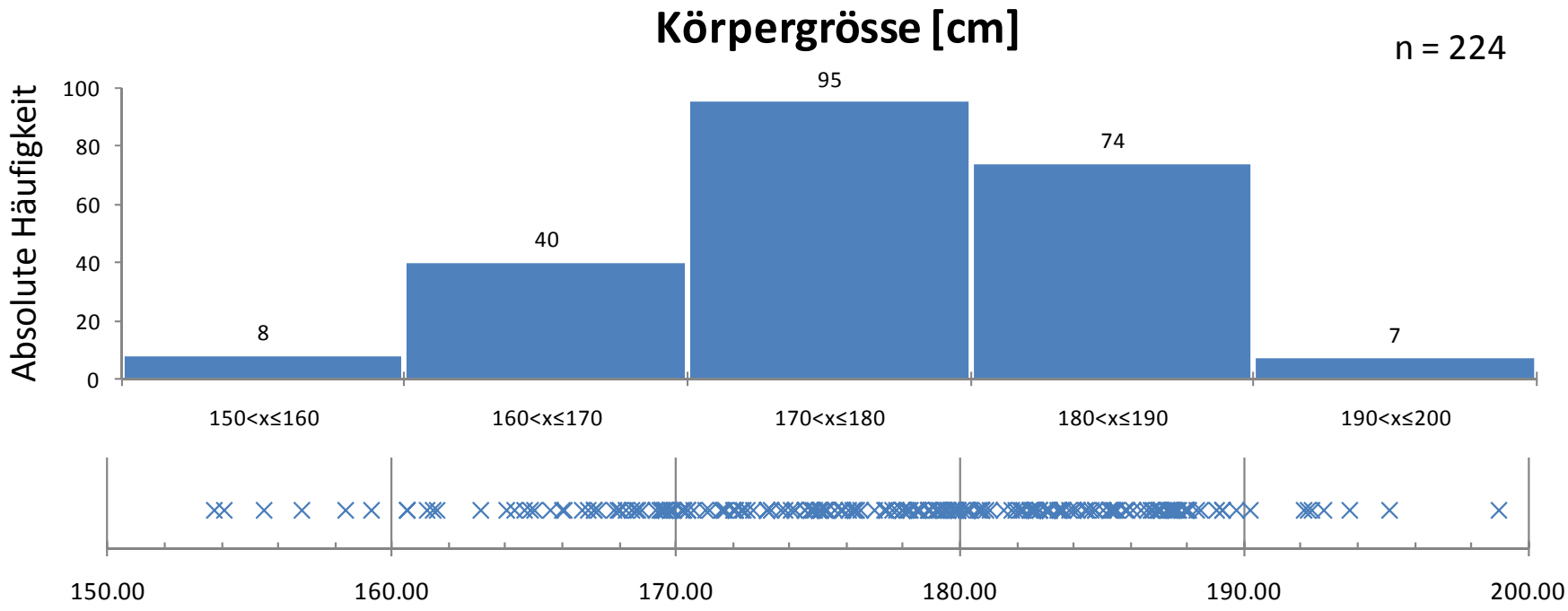
Intervall	Anzahl
150<x≤160	8
160<x≤170	40
170<x≤180	95
180<x≤190	74
190<x≤200	7
n =	224

Datenbeschreibung

Intervall	Anzahl
150<x≤160	8
160<x≤170	40
170<x≤180	95
180<x≤190	74
190<x≤200	7
n =	224

- Einfache grafische Darstellung von Stichproben

  Histogramm:



Datenbeschreibung

- Neben dem Mittelwert gibt es noch andere sog. Lageparameter:
  - Der **Median** oder Zentralwert $\tilde{x}$ der Stichprobe ist der mittlere Wert einer nach der Grösse geordneten Stichprobe $x_1^o \leq x_2^o \leq \ldots \leq x_n^o$.

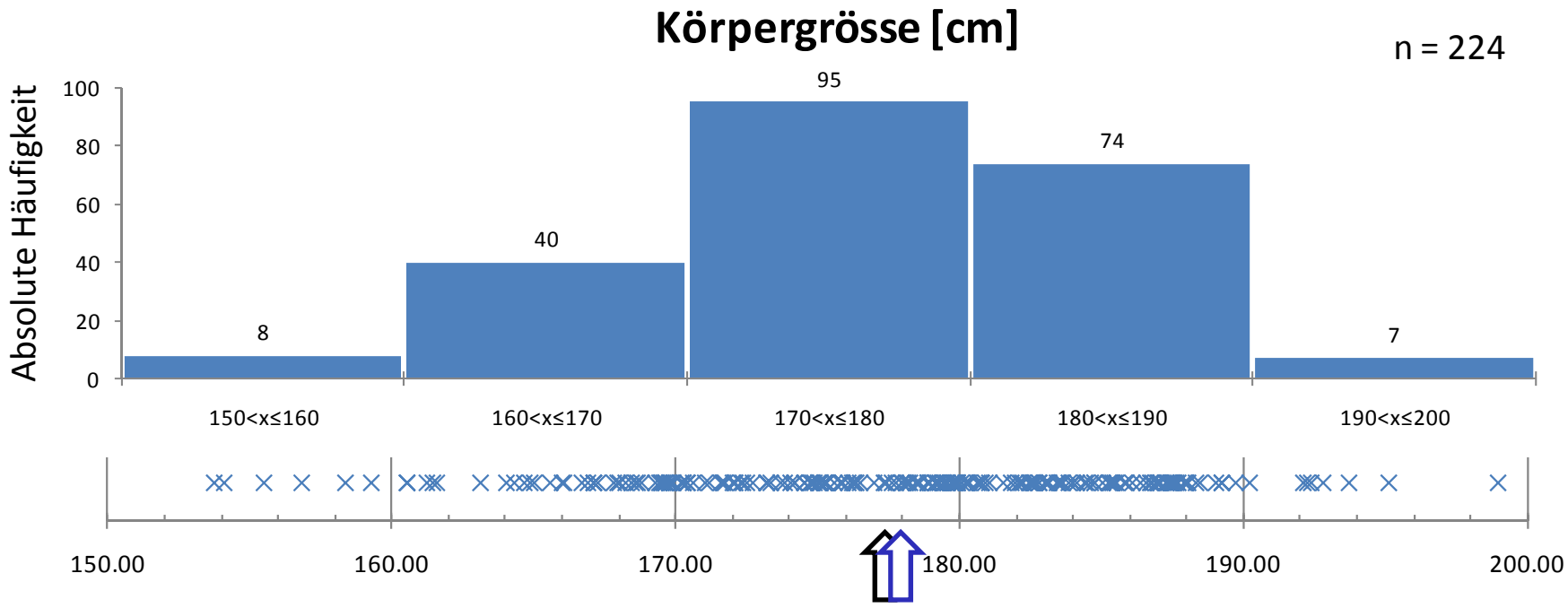
$$\tilde{x} = \begin{cases} x_{\frac{n+1}{2}} & \text{n ungerade} \\ \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1} \right) & \text{n gerade} \end{cases}$$

  - Beispiele: [23 30 31 33 120]

    [23 30 31 33]

Datenbeschreibung

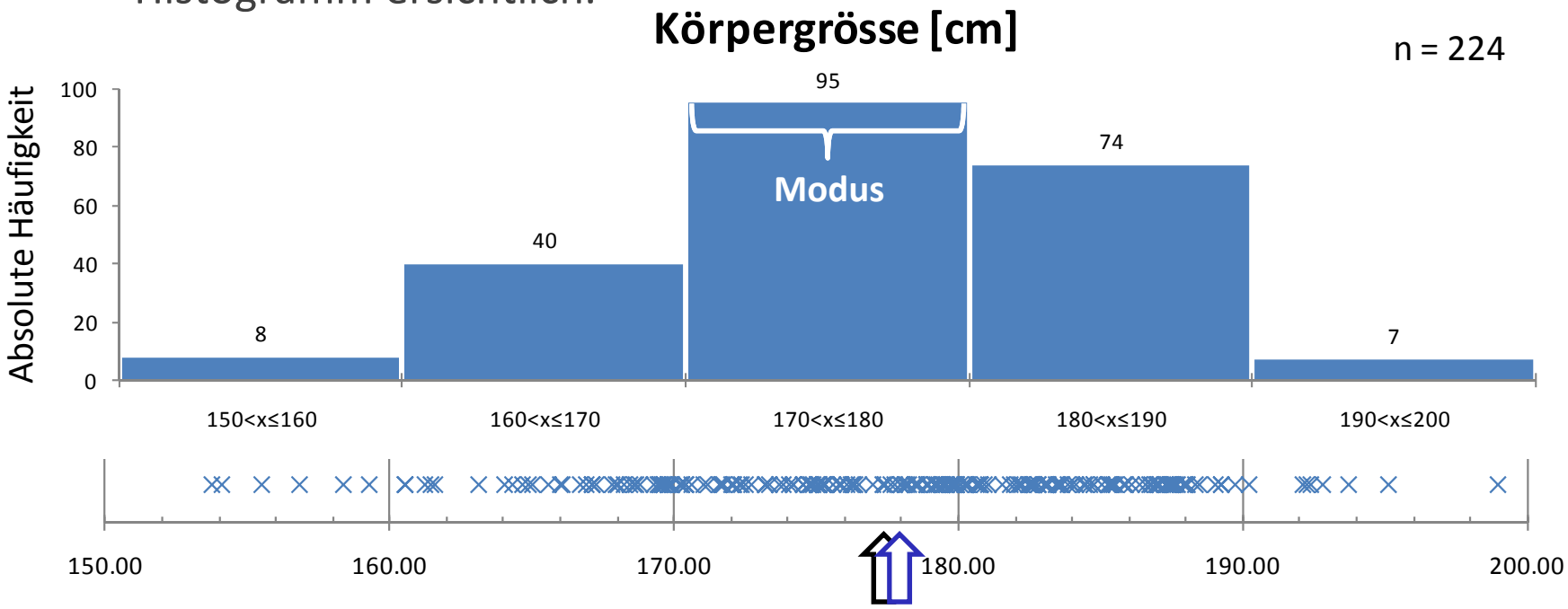
- Neben dem Mittelwert gibt es noch andere sog. Lageparameter:
  - Der **Median** oder Zentralwert $\tilde{x}$ der Stichprobe ist der mittlere Wert einer nach der Grösse geordneten Stichprobe $x_1^o \leq x_2^o \leq \ldots \leq x_n^o$.



Mittelwert = 177.64 **Median = 178.00**

Datenbeschreibung

- Neben dem Stichproben-Mittelwert gibt es noch andere sog. Lageparameter:
  - Der **Modus** oder Modalwert der Stichprobe ist der am häufigsten auftretende Wert – bei kontinuierlichen Wertemengen u.a. aus Histogramm ersichtlich.



Mittelwert = 177.64 **Median = 178.00**

Datenbeschreibung

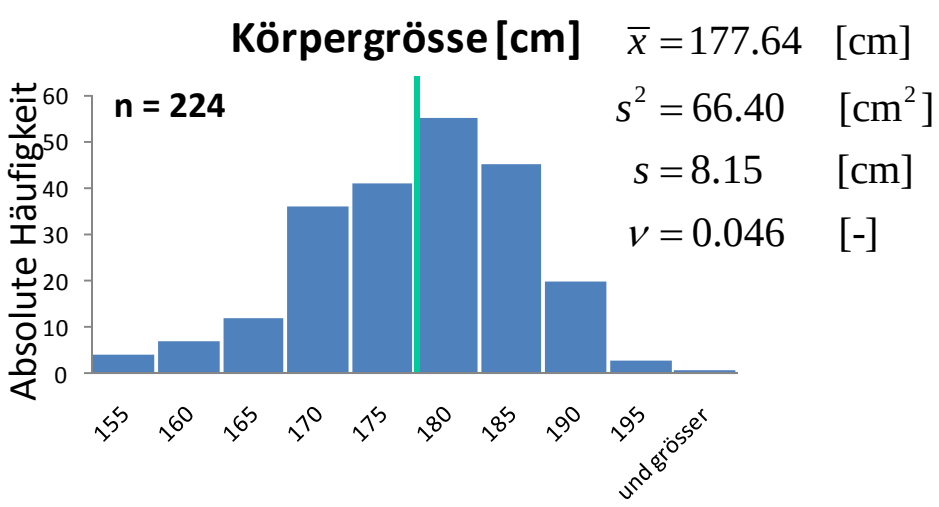
- Streuungsparameter – Streuung um den Mittelwert
  - Die Varianz der Stichprobe $$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$
  - Die Standardabweichung der Stichprobe $$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$
  - Der Variationskoeffizient der Stichprobe (relative Streuung, COV) $$\nu = \frac{s}{\bar{x}}$$

Datenbeschreibung

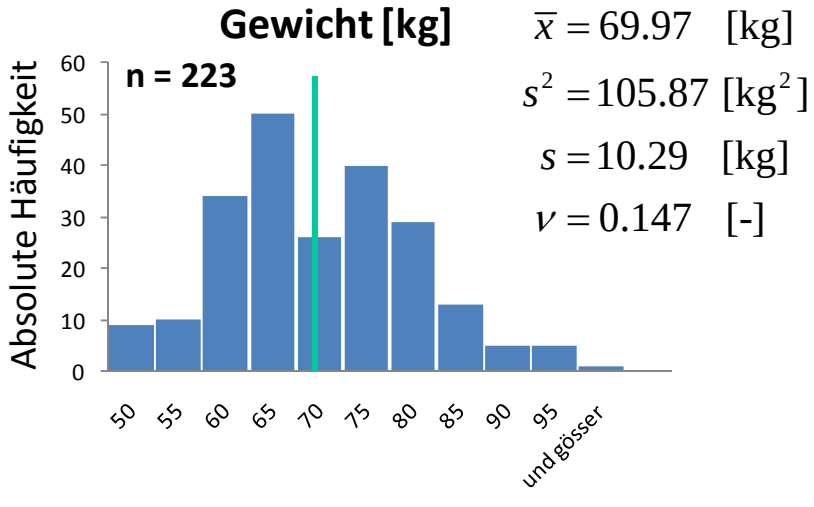
- Streuungsparameter – Streuung um den Mittelwert

Varianz $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ Standardabweichung $s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$ COV $\nu = \frac{s}{\bar{x}}$

Beispiel



$\bar{x} = 177.64$ [cm]
$s^2 = 66.40$ [cm^2]
$s = 8.15$ [cm]
$\nu = 0.046$ [-]



$\bar{x} = 69.97$ [kg]
$s^2 = 105.87$ [kg^2]
$s = 10.29$ [kg]
$\nu = 0.147$ [-]

Datenbeschreibung

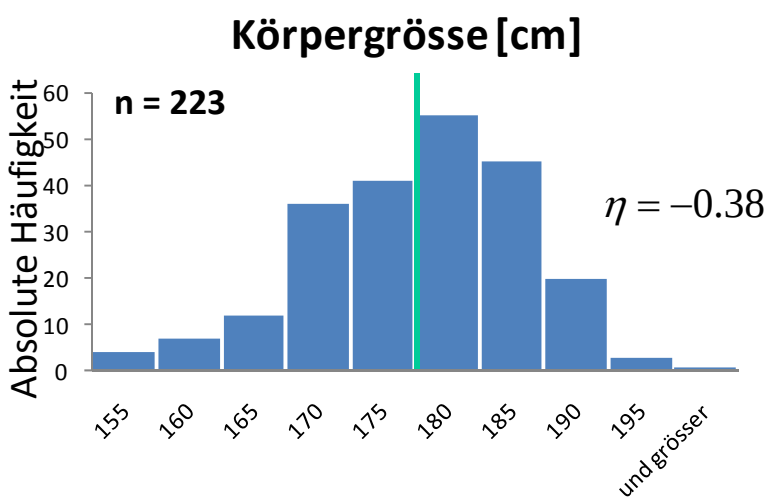
- Streuungsparameter – Streuung um den Mittelwert

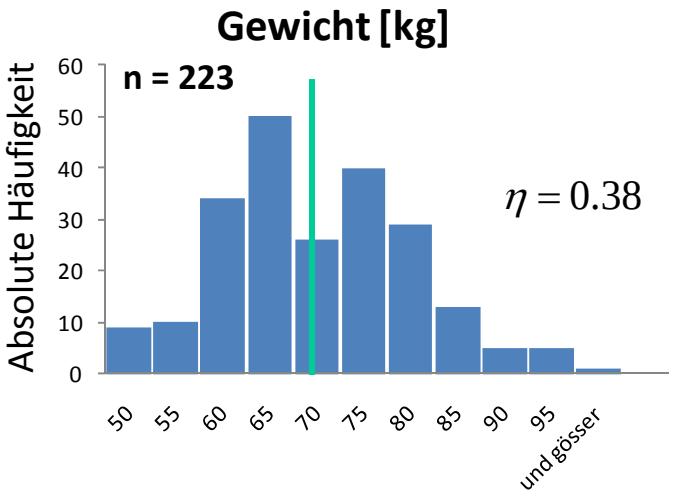
  - Der Schiefekoeffizient der Stichprobe

    -> Mass für die Asymmetrie

$$\eta = \frac{1}{n} \cdot \frac{\sum_{i=1}^{n} (x_i - \bar{x})^3}{s^3}$$

Beispiel





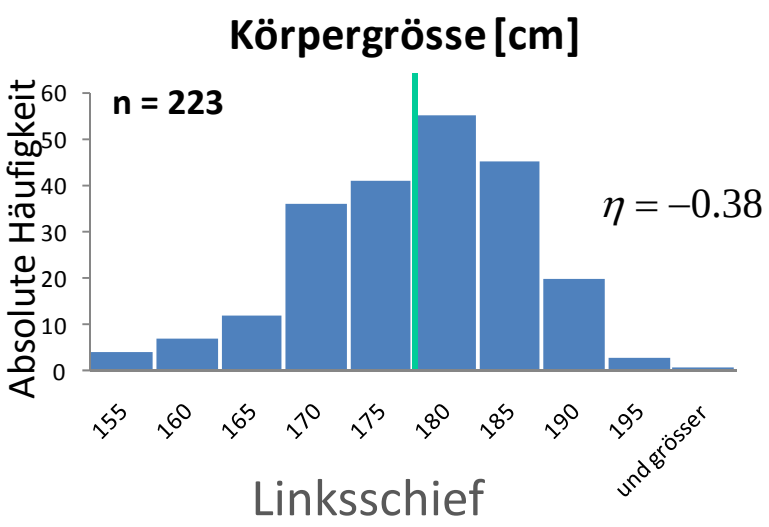
Datenbeschreibung

- Streuungsparameter – Streuung um den Mittelwert

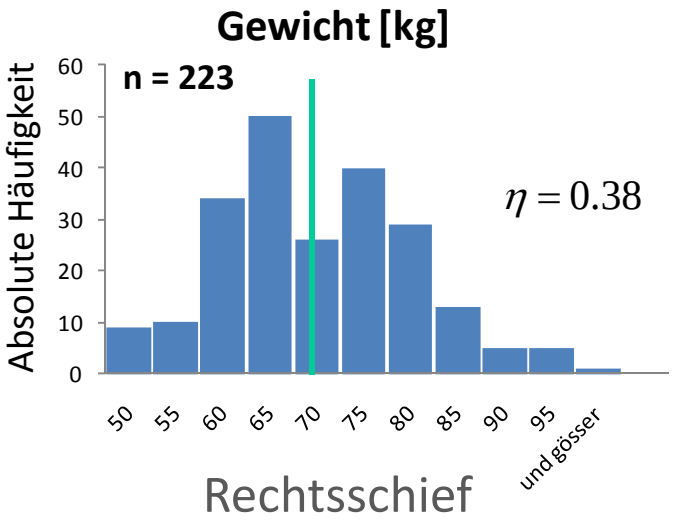
  - Der Schiefekoeffizient der Stichprobe
    -> Mass für die Asymmetrie

$$\eta = \frac{1}{n} \cdot \frac{\sum_{i=1}^{n} (x_i - \bar{x})^3}{s^3}$$

Beispiel



Linksschief



Rechtsschief

Datenbeschreibung

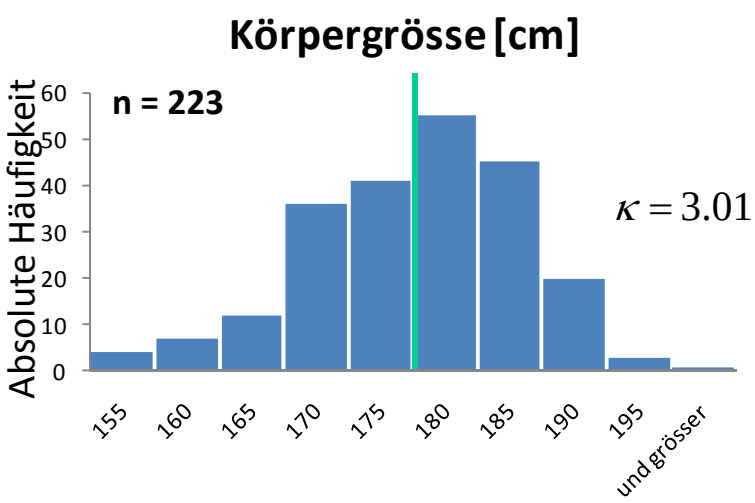
- Streuungsparameter – Streuung um den Mittelwert

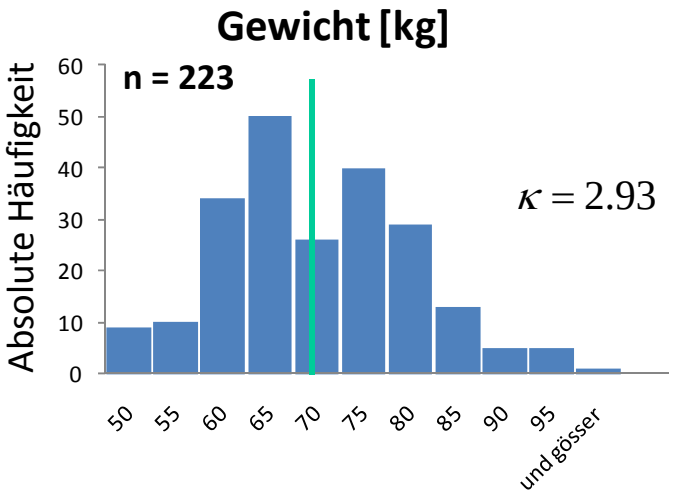
  - Kurtosis der Stichprobe:

    -> Mass für die Spitzigkeit / Gipfligkeit

$$\kappa = \frac{1}{n} \cdot \frac{\sum_{i=1}^{n} (x_i - \bar{x})^4}{s^4}$$

Beispiel





Datenbeschreibung

- Beschreibung von paarweise beobachteten Eigenschaften

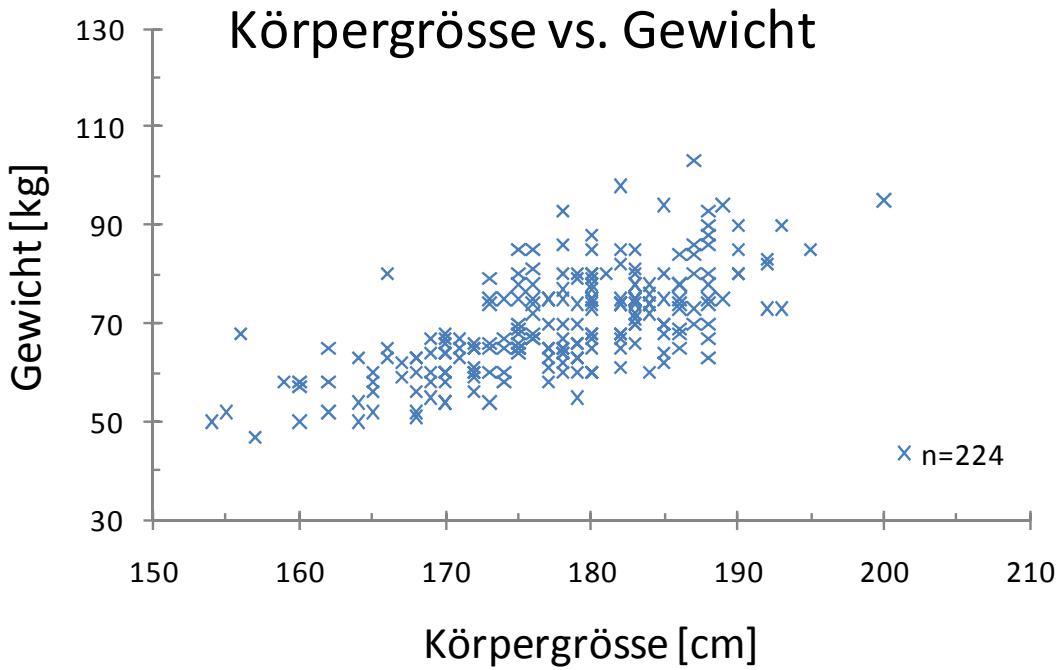
$$\mathbf{x} = \left(x_1, x_2, x_3, \ldots, x_n \right)^T$$

$$\mathbf{y} = \left(y_1, y_2, y_3, \ldots, y_n \right)^T$$

Datenbeschreibung

- Beschreibung von paarweise beobachteten Eigenschaften

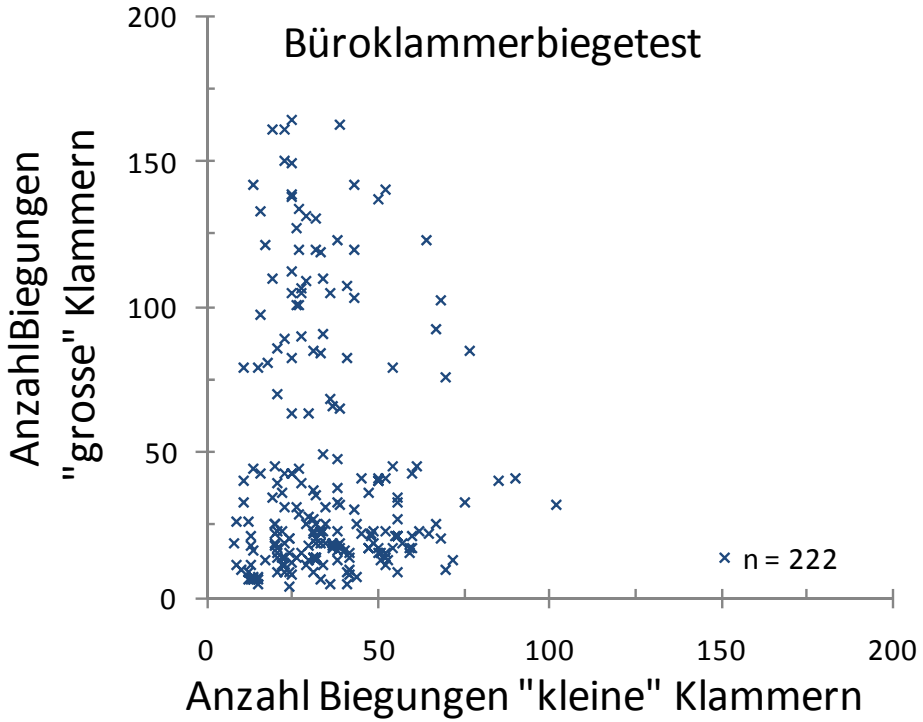
  Das zweidimensionale Streudiagramm



Datenbeschreibung

- Beschreibung von paarweise beobachteten Eigenschaften

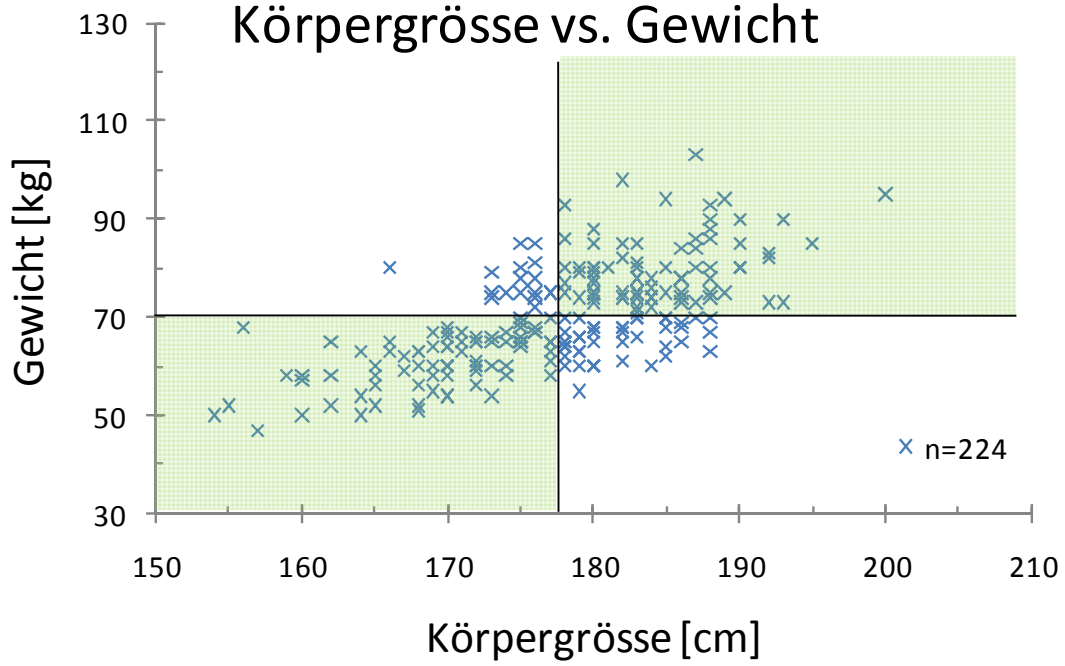
  Das zweidimensionale Streudiagramm



Datenbeschreibung

- Beschreibung von paarweise beobachteten Eigenschaften

- Die Kovarianz der Stichprobe: $s_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})$



$x \rightarrow$ Körpergrösse

$\bar{x} = 177.64$ cm

$y \rightarrow$ Gewicht

$\bar{y} = 69.97$ kg

Datenbeschreibung

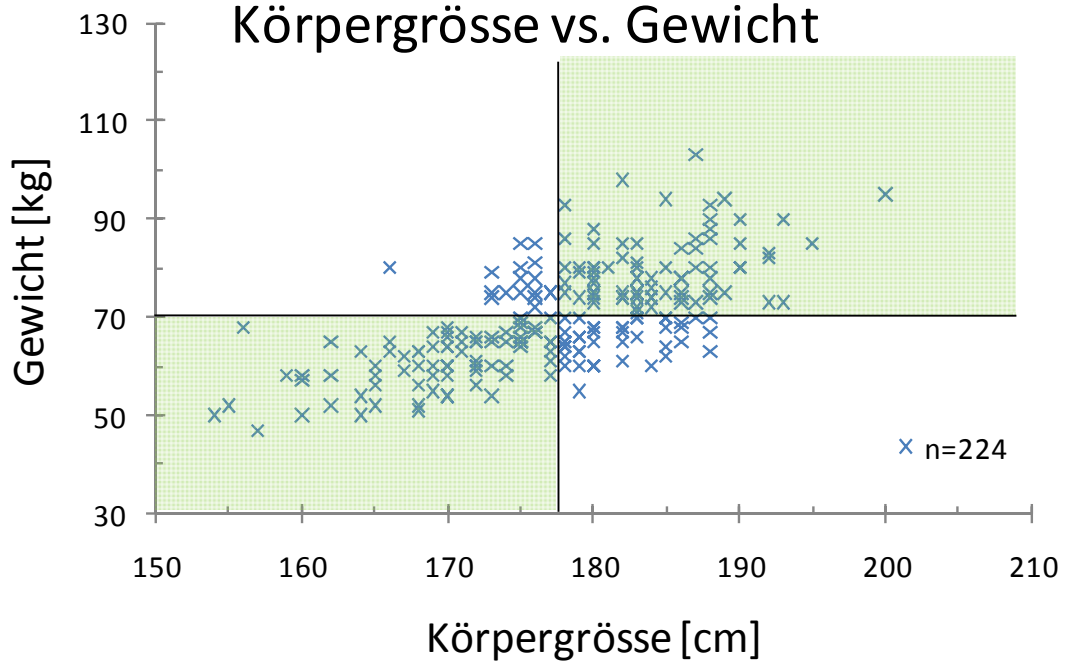
- Beschreibung von paarweise beobachteten Eigenschaften
- Die Kovarianz der Stichprobe: $s_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})$
- Der Korrelationskoeffizient der Stichprobe:

$$r_{XY} = \frac{1}{n} \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{s_X \cdot s_Y}$$

ist limitiert auf das Interval $[-1,1]$

Datenbeschreibung

- Beschreibung von paarweise beobachteten Eigenschaften
- Der Korrelationskoeffizient: $r_{XY} = \frac{1}{n} \frac{\sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{s_X \cdot s_Y} = 0.6762$



$x \rightarrow$ Körpergrösse

$\bar{x} = 177.64$ cm

$y \rightarrow$ Gewicht

$\bar{y} = 69.97$ kg

Nummerische Kennwerte

Lageparameter:

Arithmetisches Mittel	Schwerpunkt der Stichprobe
Median	mittlerer Wert einer Stichprobe
Modalwert	am häufigsten vorkommender Wert

Streuungsparameter:

Varianz / Standardabweichung	Verteilung um den Mittelwert
Variationskoeffizient	Variabilität relativ zum Mittelwert

Andere Parameter:

Schiefekoeffizient	Schiefe relativ zum Mittelwert
Kurtosis	Spitzigkeit/Gipfligkeit um den Mittelwert

Masse für Korrelation:

Kovarianz	Tendenz für paarweise beobachtete Eigenschaften
Korrelationskoeffizient	Normalisierter Koeffizient zwischen -1 und +1

Weitere grafische Darstellungsformen

- Histogramm Fortsetzung
- Quantil-Plots
- Tukey Box Plots

Histogramm

- Prinzip:
  - Aufteilung der Stichprobe in *k* Grössen-Intervalle
  - Auftragen der Häufigkeit je Intervall

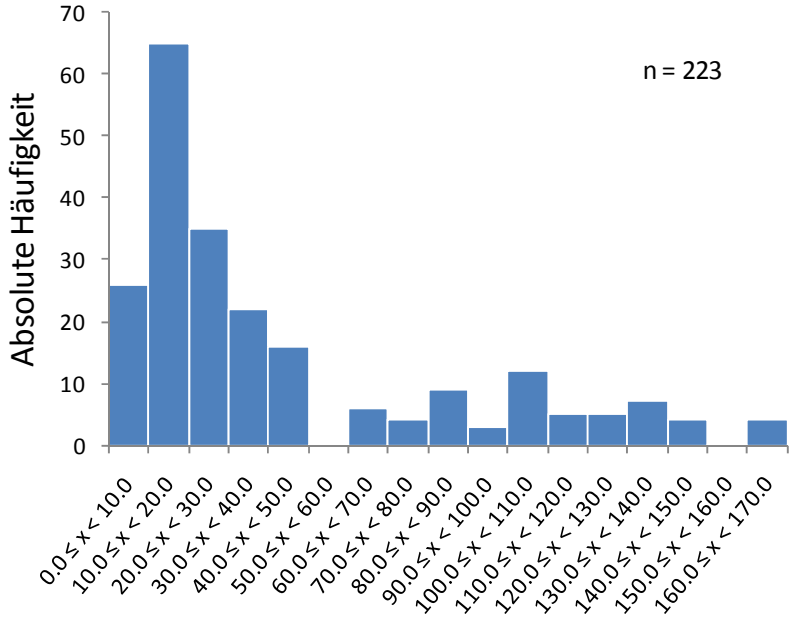
- Beispiel: Ihre Büroklammerdaten vom letzten Mal
  „grosse“ Klammern, Stichprobenumfang n = 223,
  Maximalwert 164, Minimalwert 4.

  Einteilung in 17 Intervalle; [0,10); [10,20); [20,30);… ; [160,170)

Histogramm

- Prinzip:
  - Aufteilung der Stichprobe in *k* Grössen-Intervalle
  - Auftragen der Häufigkeit je Intervall

- Beispiel:

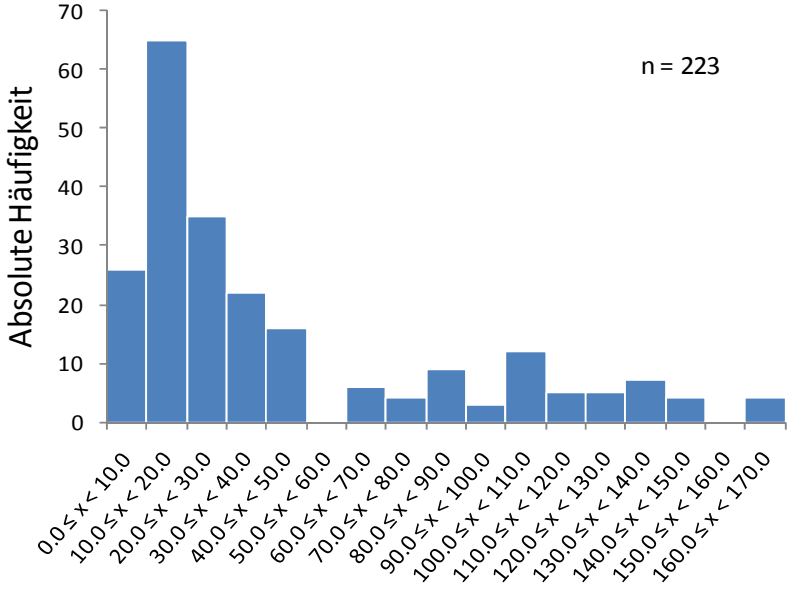


Anzahl Biegungen der „grossen“ Klammern

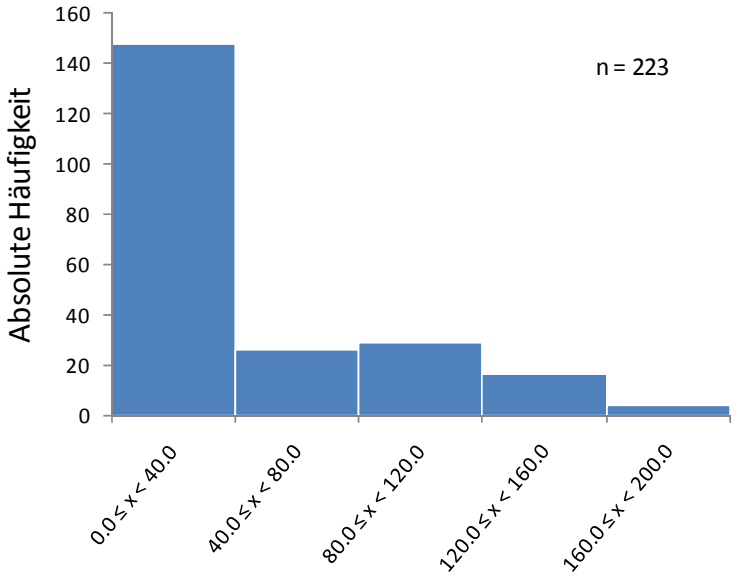
Histogramm

- Prinzip:
  - Aufteilung der Stichprobe in *k* Grössen-Intervalle
  - Auftragen der Häufigkeit je Intervall
- Beispiel:

Aussage abhängig von der Anzahl der Intervalle!



Anzahl Biegungen der „grossen“ Klammern



Anzahl Biegungen der „grossen“ Klammern

Histogramm

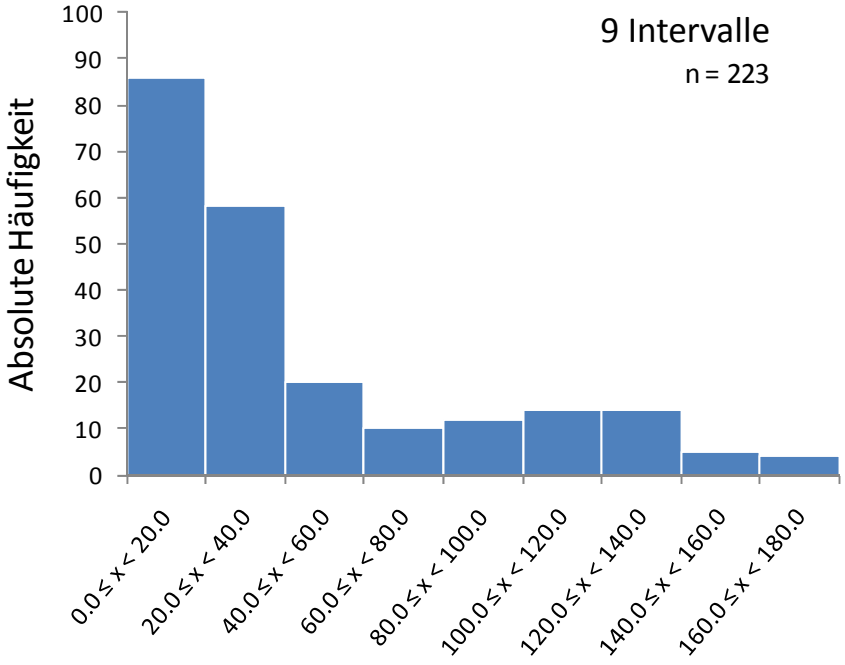
- Prinzip:
  - Aufteilung der Stichprobe in *k* Grössen-Intervalle
  - Auftragen der Häufigkeit je Intervall
  - **Faustregel für die Anzahl der Intervalle:** $k = 1 + 3.3\log(n)$
  -
- Beispiel: Büroklammerdaten „grosse“ Klammern,
  Stichprobenumfang n = 223, Wertebereich [4,164]

$$k = 1 + 3.3\log(223) = 8.75 \cong 9 \text{ Intervalle}$$

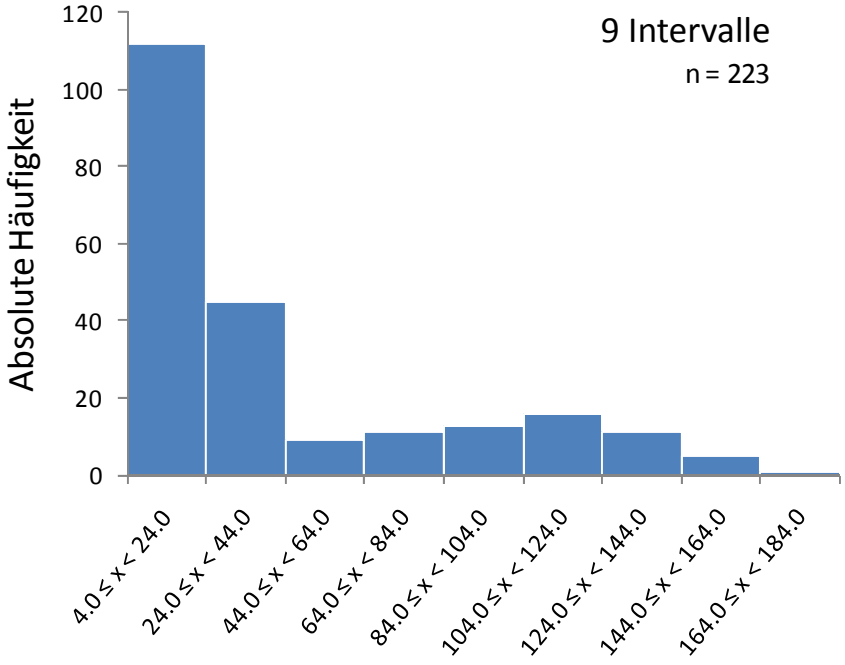
[0,20); [20,40); [40,60);… ; [160,180)

oder [4,24); [24,44); [44,64);… ; [164,184.0) ?

Histogramm



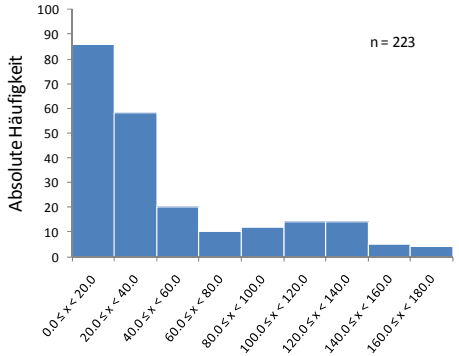
Anzahl Biegungen der „grossen“ Klammern



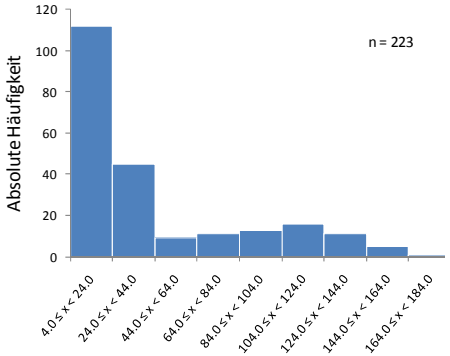
Anzahl Biegungen der „grossen“ Klammern

Histogramm

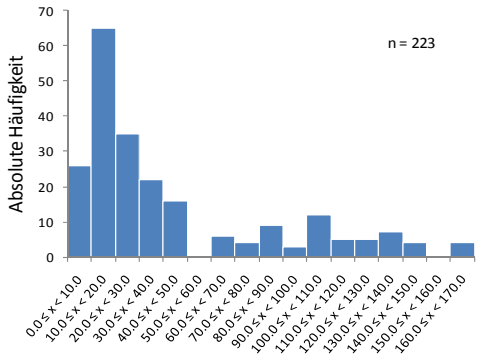
- Die Form des Histogramms hängt ab von
  - der Anzahl der Intervalle.
  - der Wahl des Startpunktes.



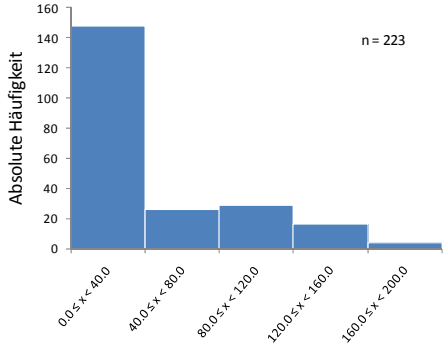
Anzahl Biegungen der „grossen“ Klammern



Anzahl Biegungen der „grossen“ Klammern



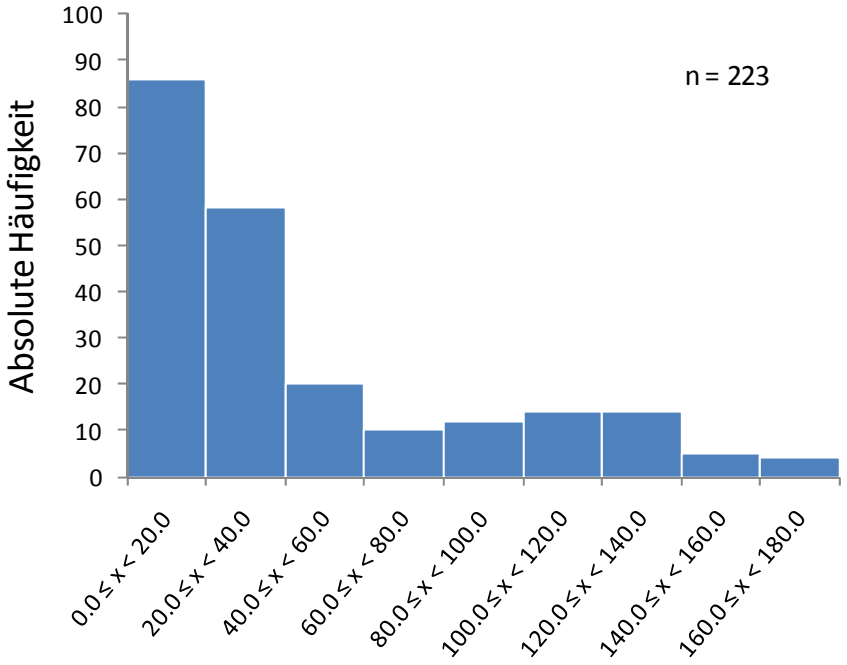
Anzahl Biegungen der „grossen“ Klammern



Anzahl Biegungen der „grossen“ Klammern

Histogramm

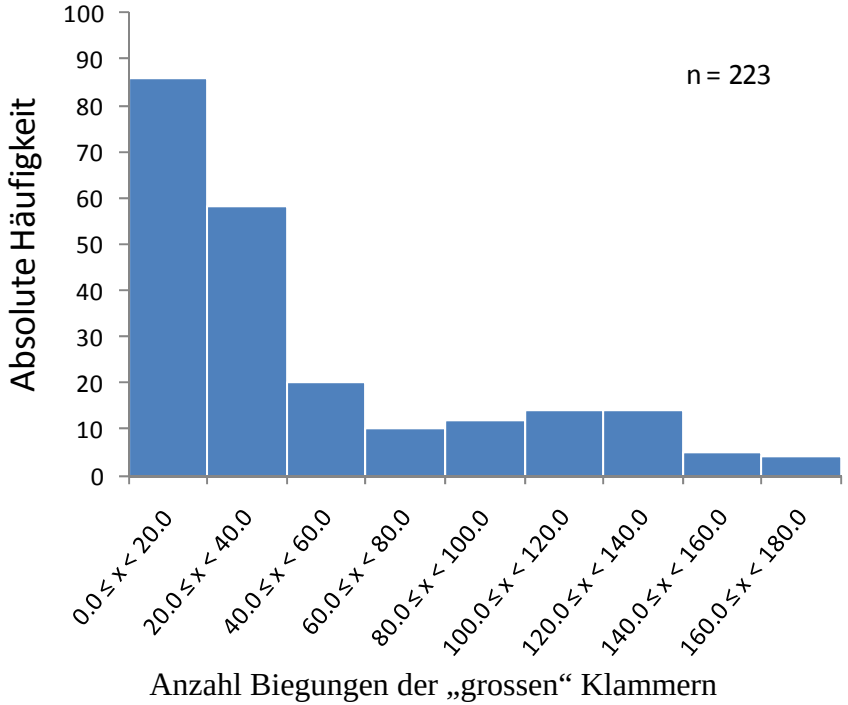
- Bisher haben wir die absolute Häufigkeit betrachtet.

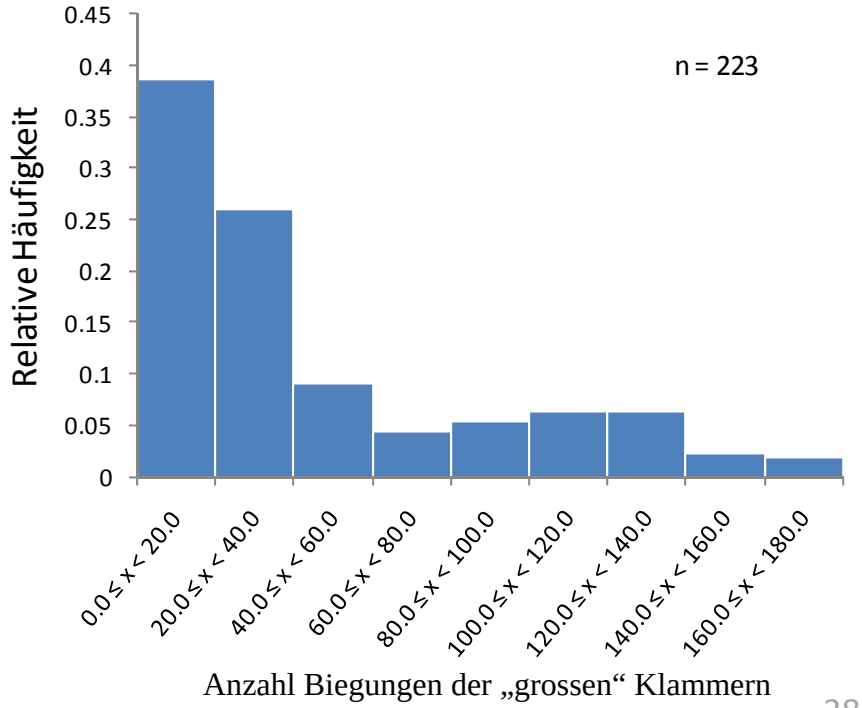


Anzahl Biegungen der „grossen“ Klammern

Histogramm

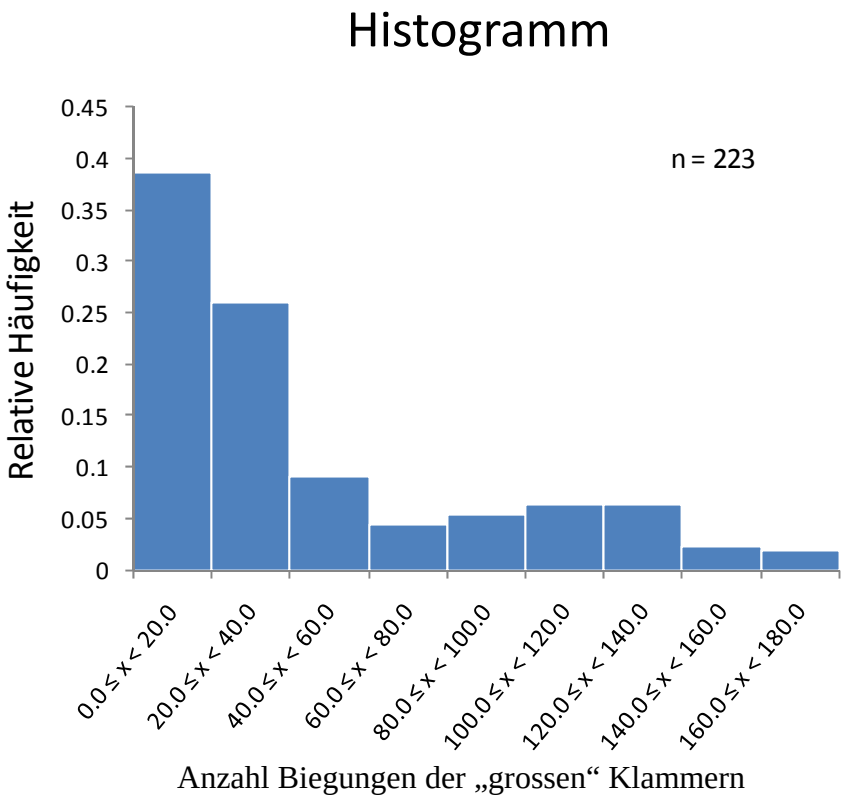
- Bisher haben wir die absolute Häufigkeit betrachtet.
- In der Regel wird die Häufigkeit relativ, also normiert betrachtet.

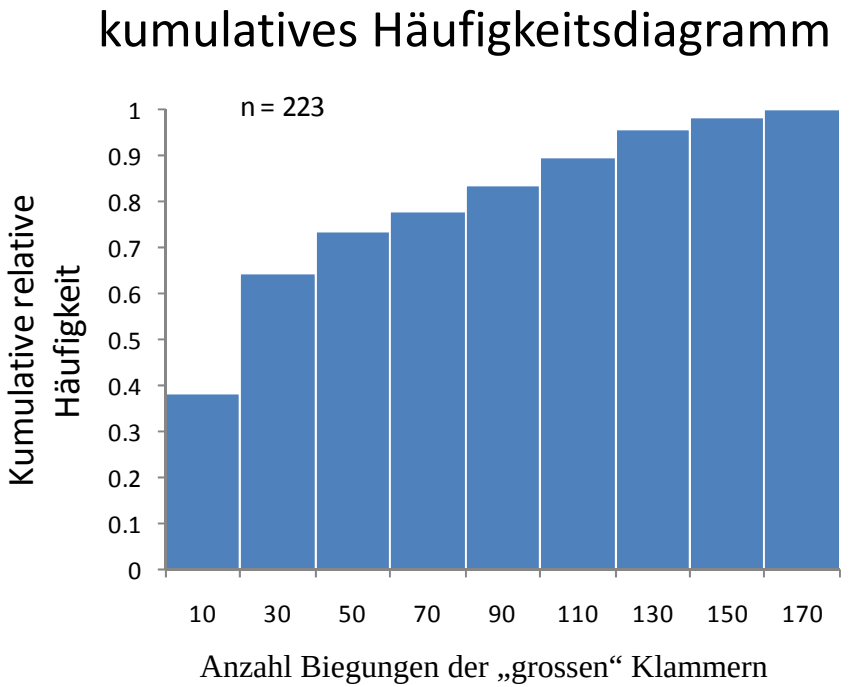




Histogramm

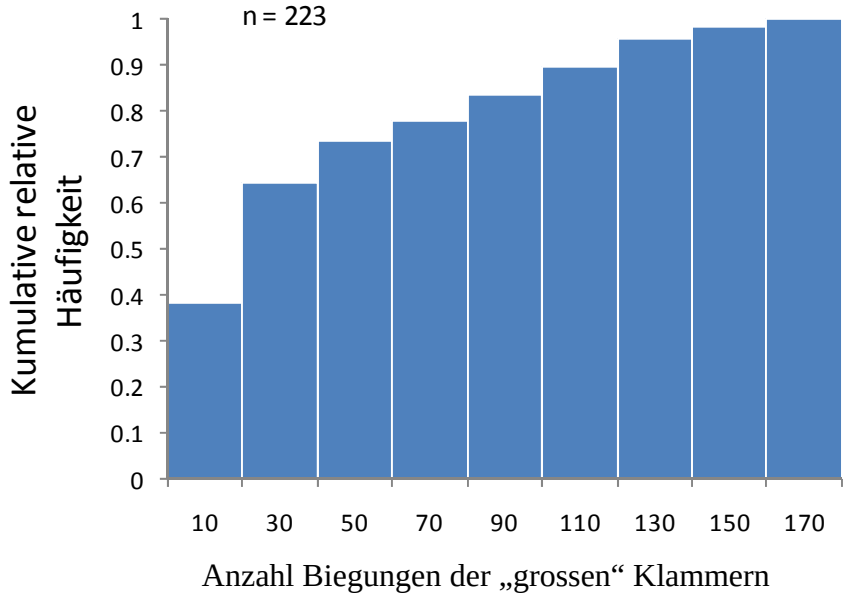
- Eine Spielart des Histogramms ist das kumulative Häufigkeitsdiagramm.

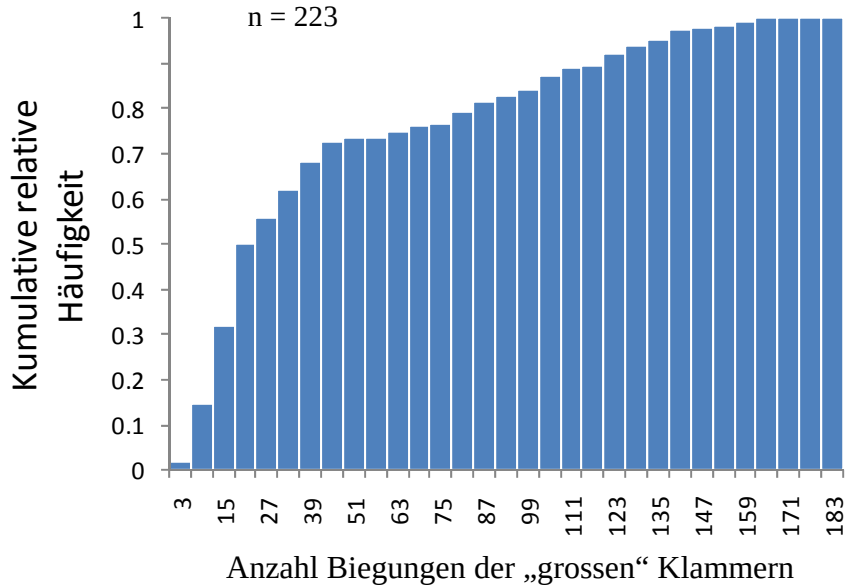




Histogramm

- Eine Spielart des Histogramms ist das kumulative Häufigkeitsdiagramm.
- Hier kann die Intervalleinteilung beliebig klein sein!





Weitere grafische Darstellungsformen

- Histogramm Teil II.
- Quantil-Plots
- Tukey Box Plots

Quantil - Plot

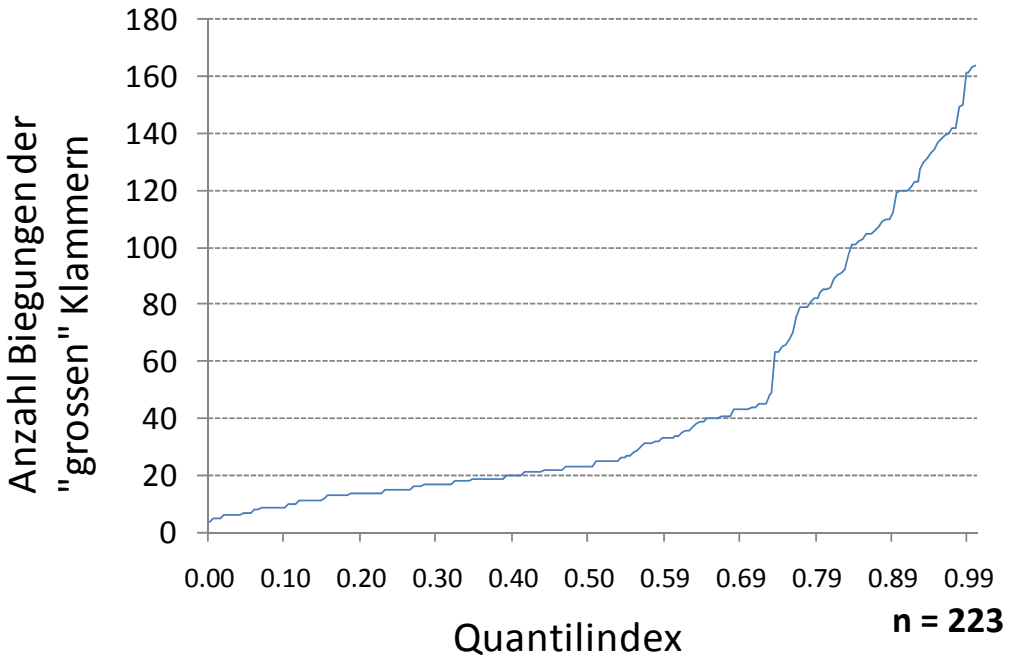
Das Quantil ist für eine gegebene Anzahl an Beobachtungen wie folgt definiert:

- Das υ-Quantil ist der Wert, der die unteren $\upsilon \cdot 100\%$ der Messwerte von den oberen $100\% - \upsilon \cdot 100\%$ trennt.
- Beispiel: Das 0.75-Quantil wird von $100\% - 0.75 \cdot 100\% = 25\%$ der Daten überschritten.
- Die Quantile werden von der geordneten (sortierten) Stichprobe berechnet: $x_1^o \le x_2^o \le \ldots \le x_n^o$
- Der Quantilindex wird wie folgt berechnet:

$$\upsilon = \frac{i}{n+1}; \quad n: \text{ Gesamt Anzahl der Beobachtungen, Rang } i=1,2\ldots,n$$

Quantil - Plot

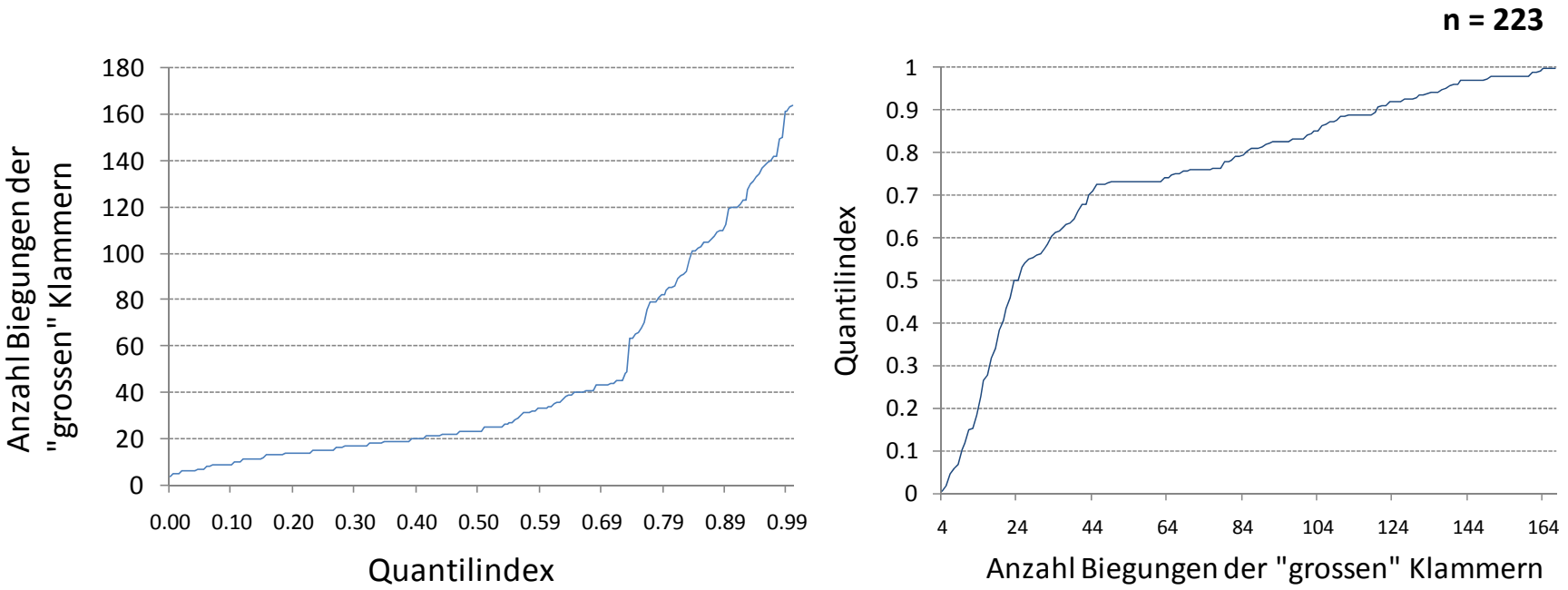
- Quantil-Plots werden durch Auftragen der Daten und der Quantilindizes gebildet.



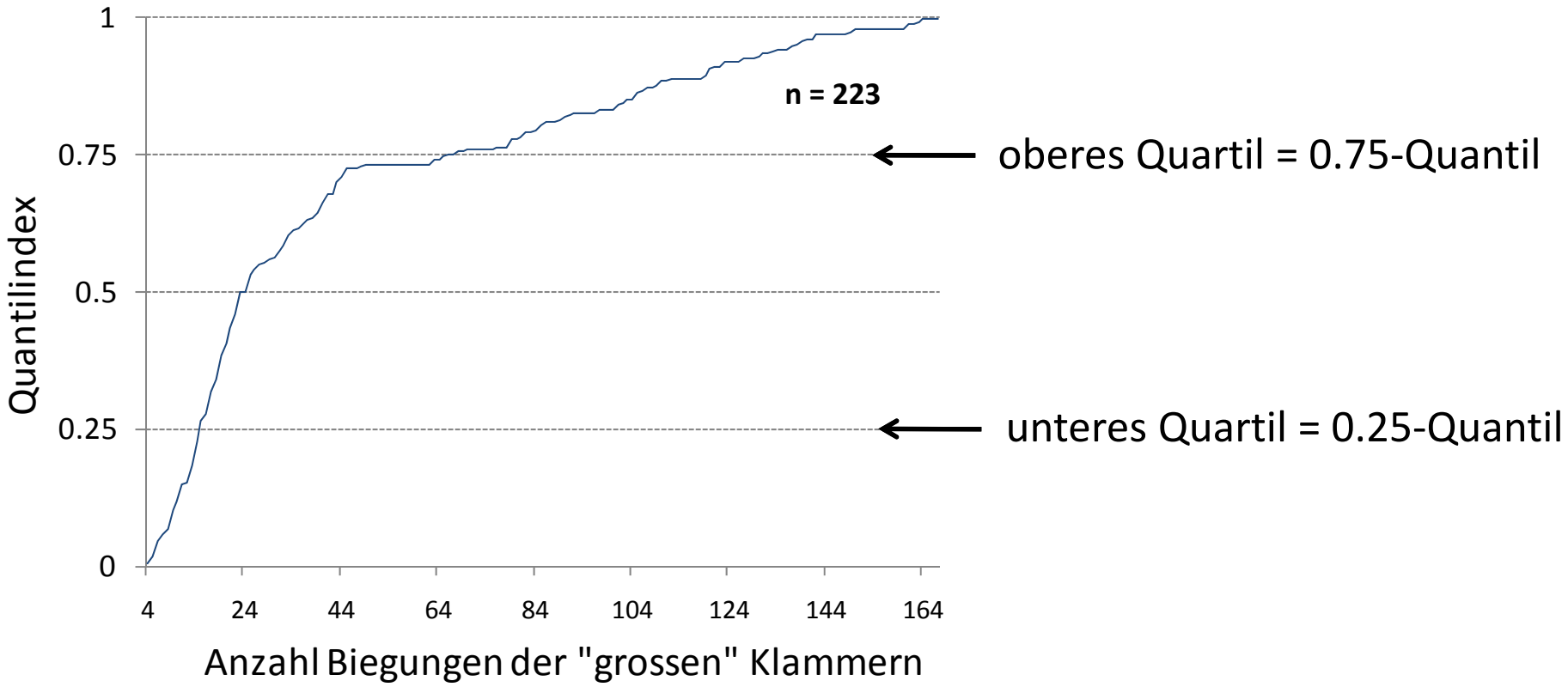
i	$i/n+1$	x_i
1	0.004878	6
2	0.009756	8
3	0.014634	9
4	0.019512	10
5	0.02439	10
6	0.029268	10
7	0.034146	11
8	0.039024	12
9	0.043902	12

Kumulatives Häufigkeitsdiagramm

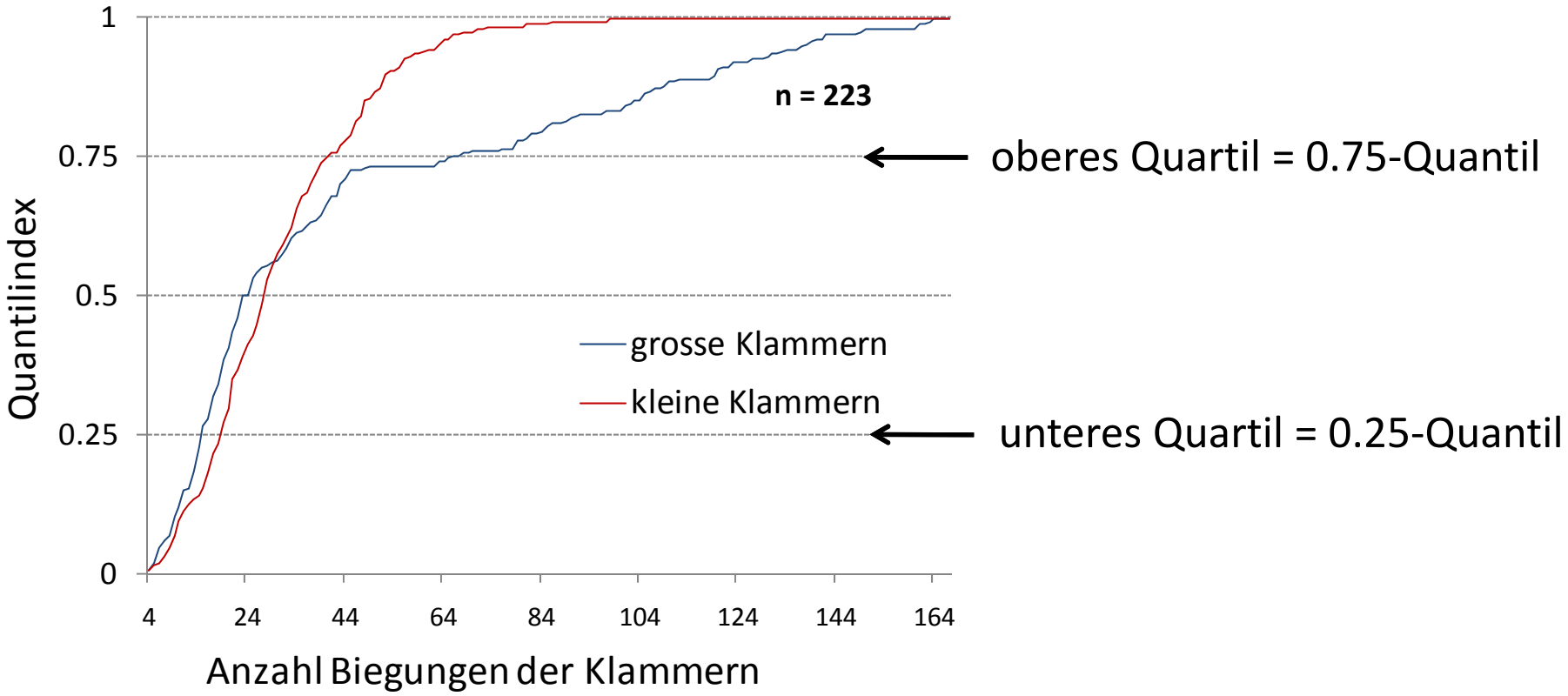
- Werden im Quantil-Plots die Achsen vertauscht ergibt dies ein kumulatives Häufigkeitsdiagramm.



Kumulatives Häufigkeitsdiagramm



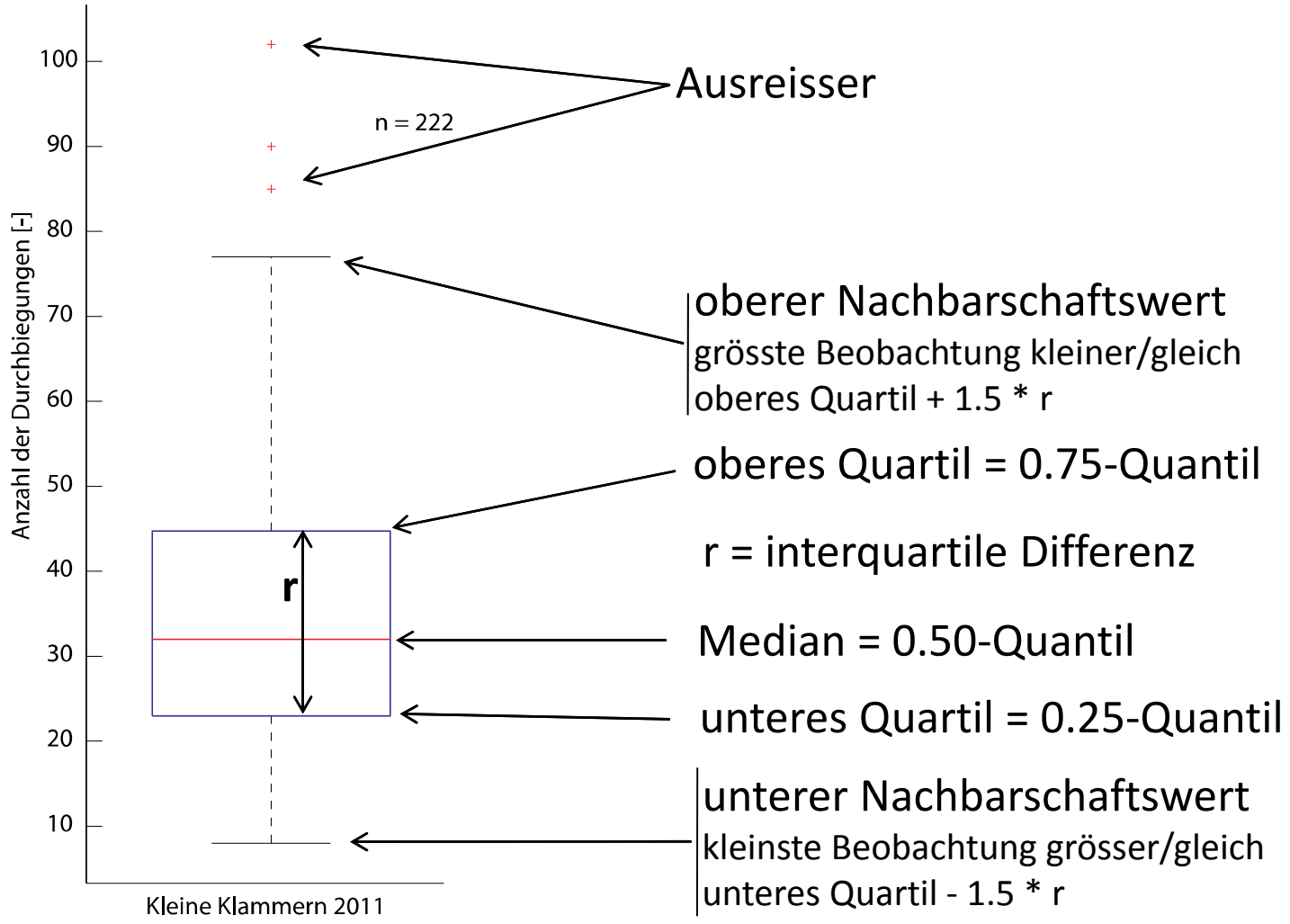
Kumulatives Häufigkeitsdiagramm



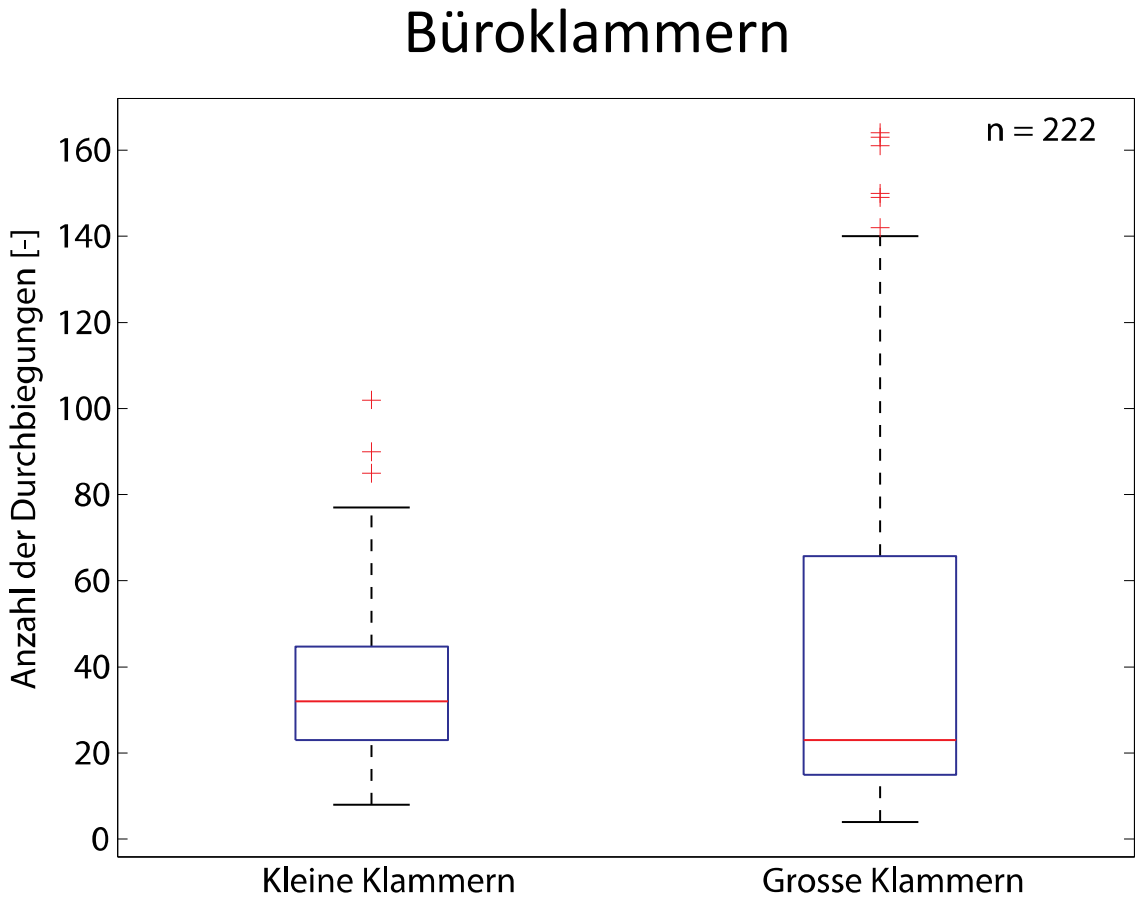
Tukey Box Plot

- Der Tukey Box Plot illustriert:
  - Median
  - untere und obere Quartilwerte
  - unterer und oberer Nachbarschaftswert
  - interquartile Differenz
  - Ausreisser

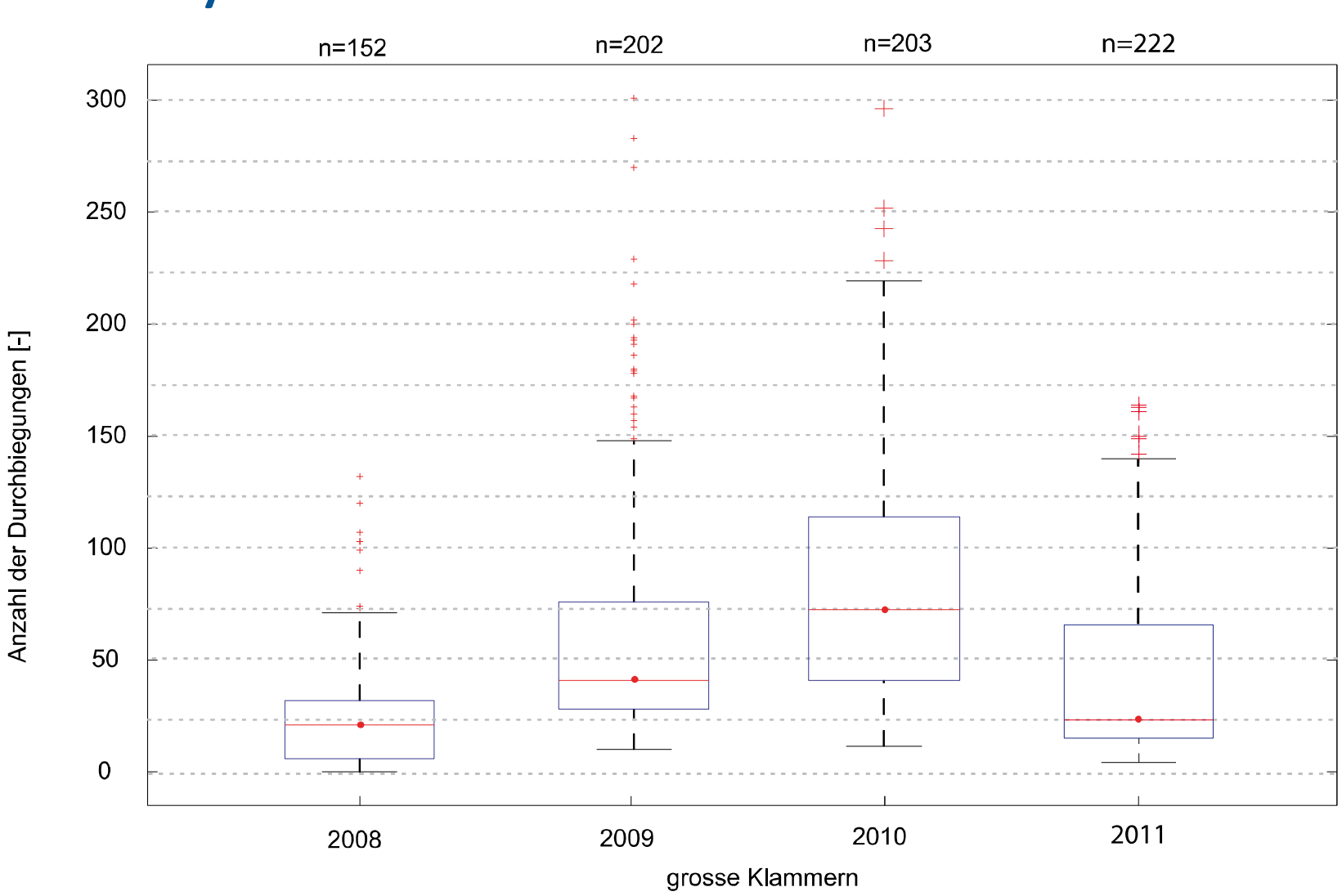
Tukey Box Plot



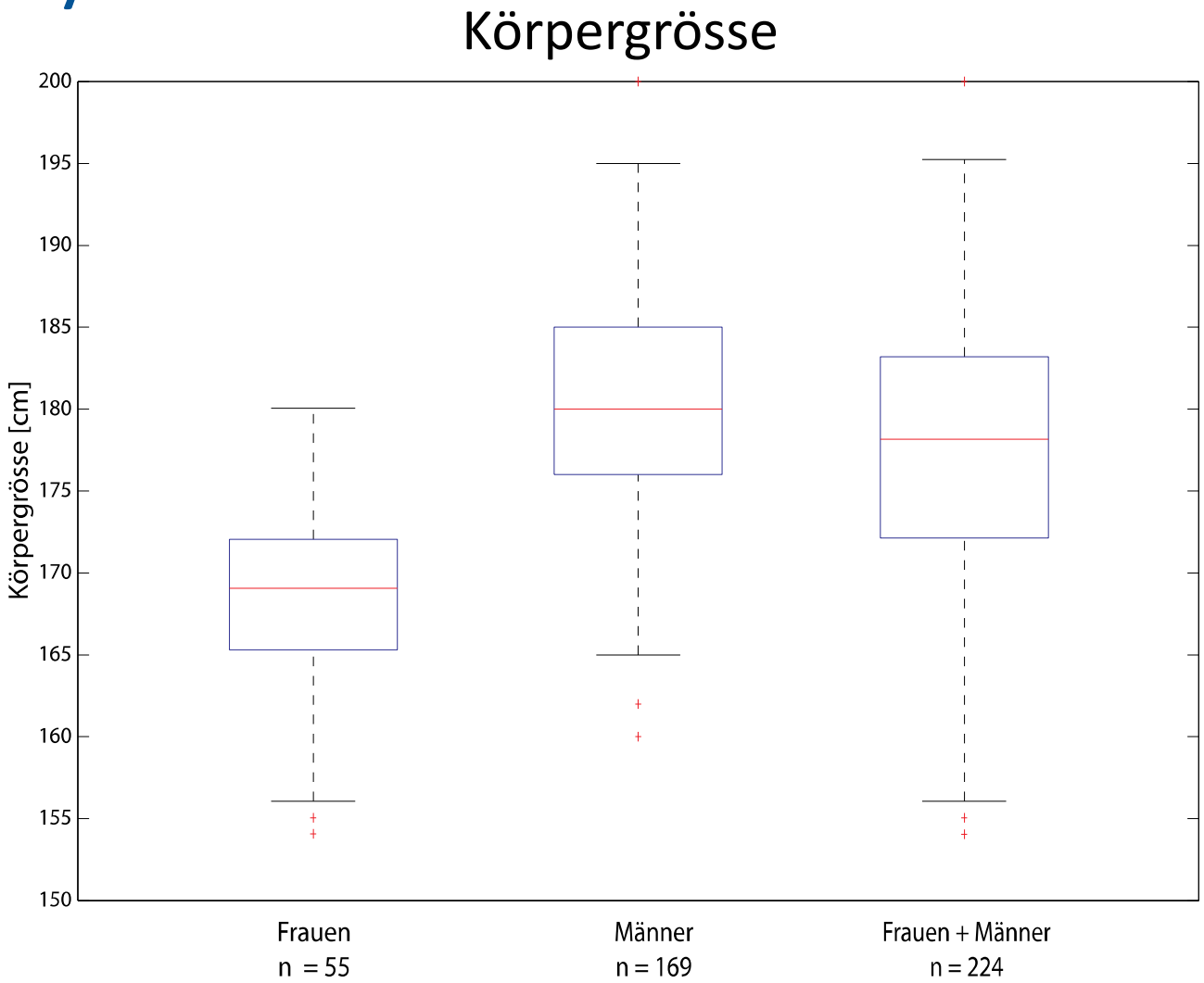
Tukey Box Plot



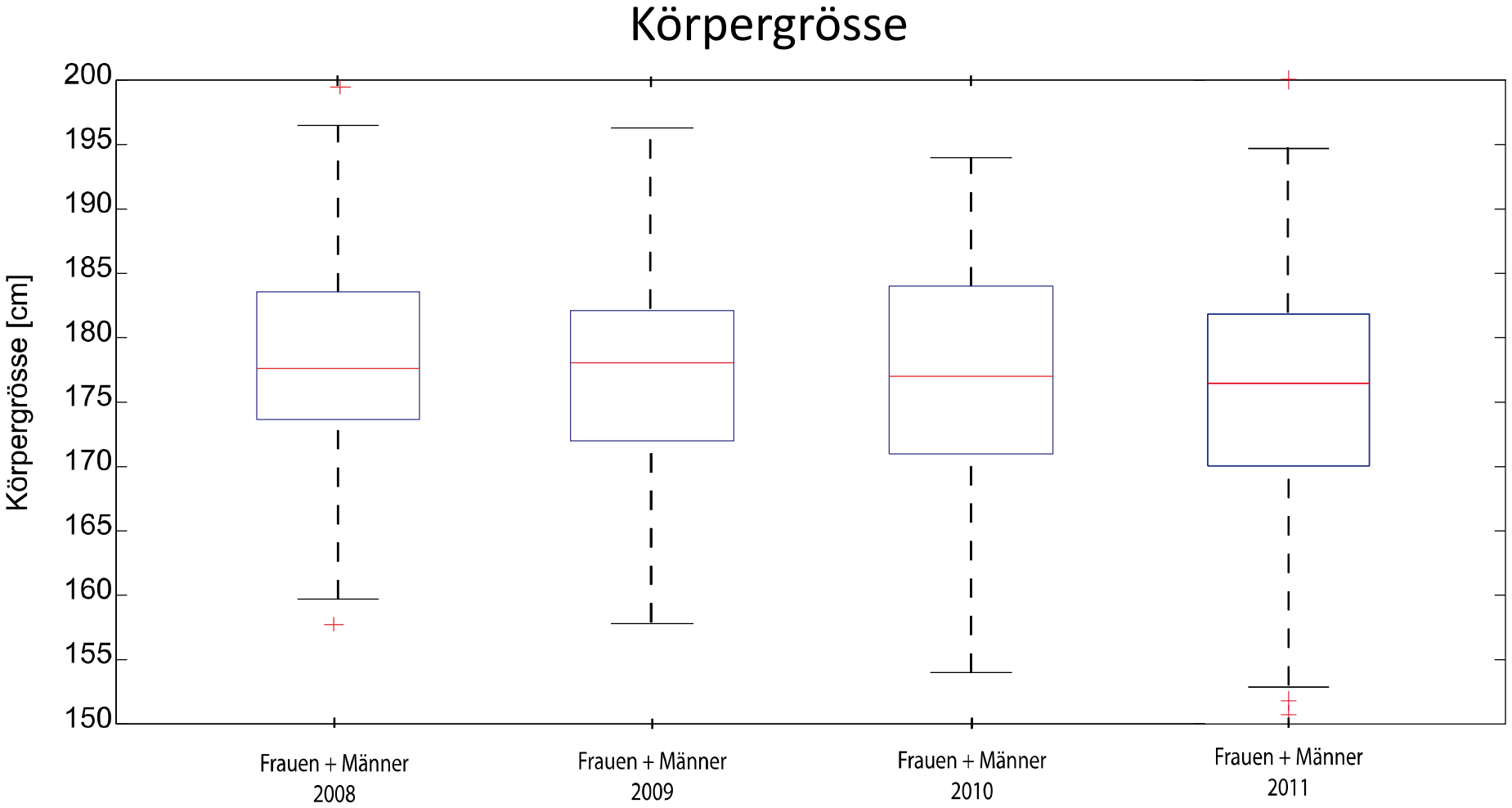
Tukey Box Plot



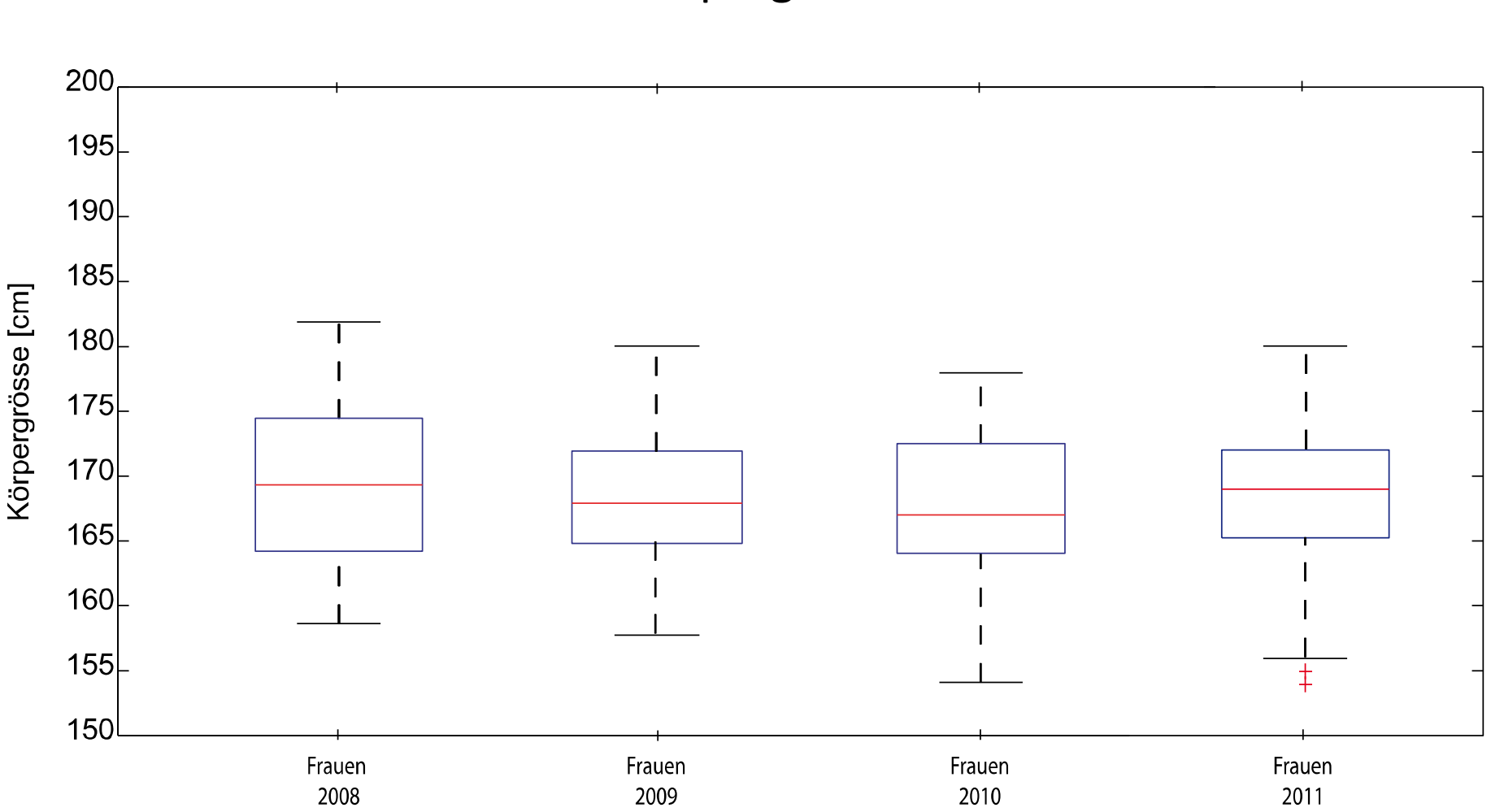
Tukey Box Plot



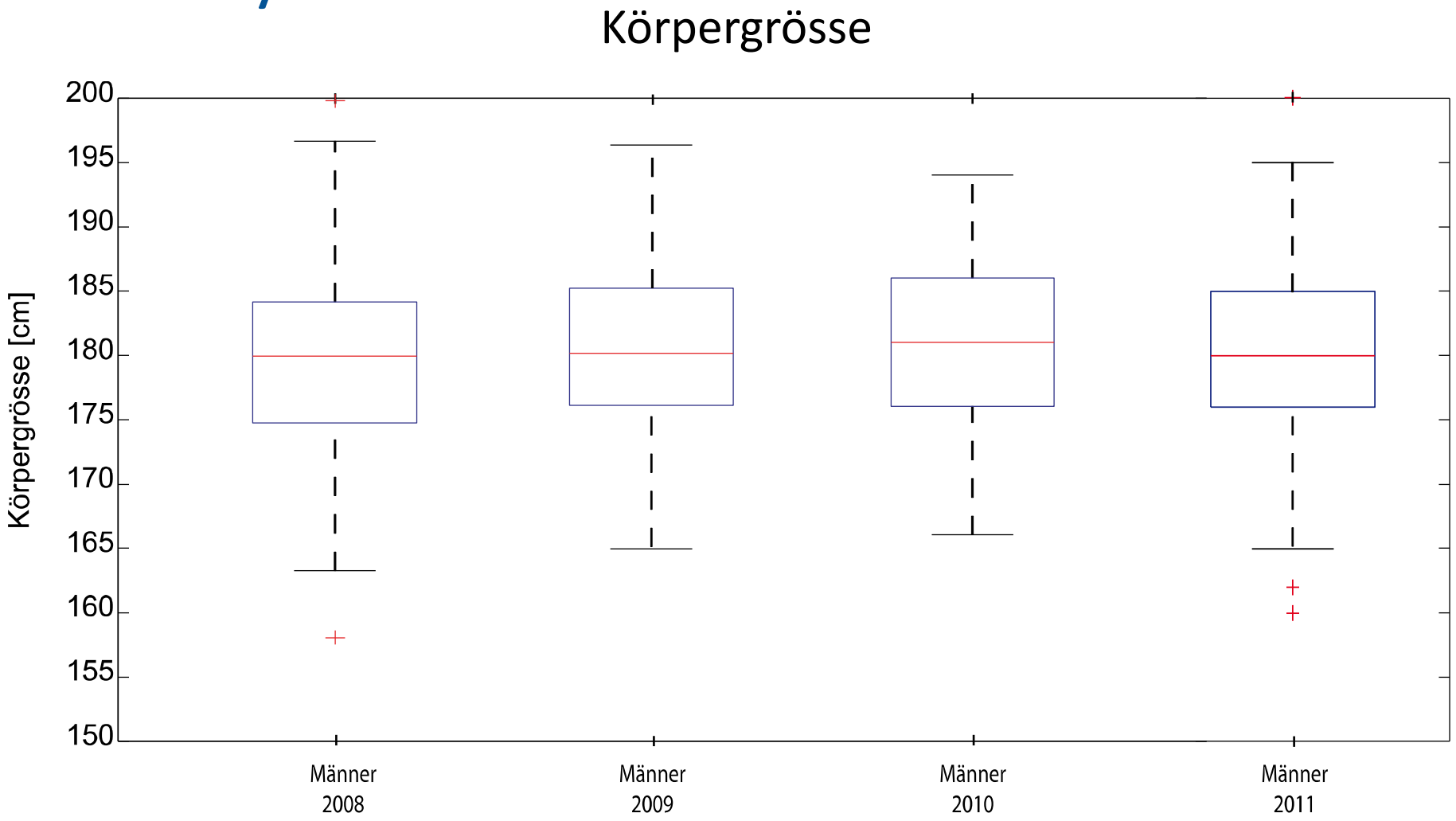
Tukey Box Plot



Tukey Box Plot

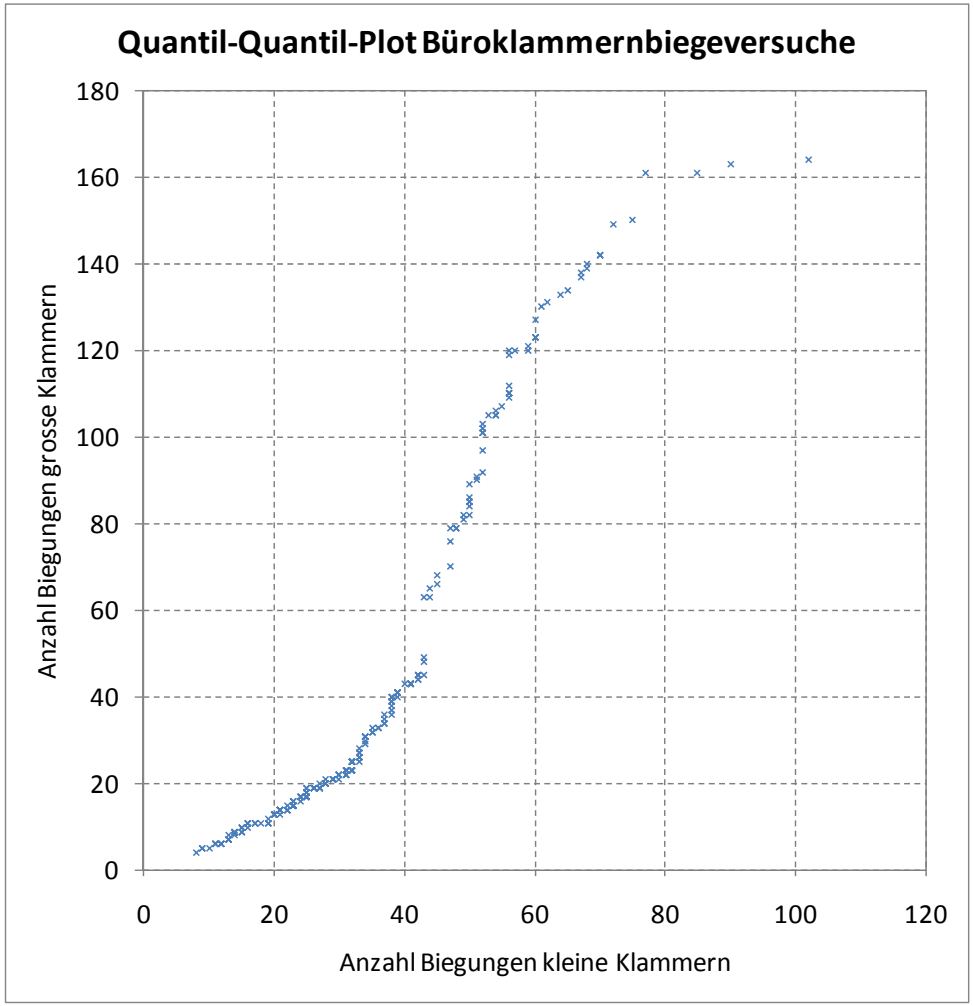


Tukey Box Plot



Q-Q Plots

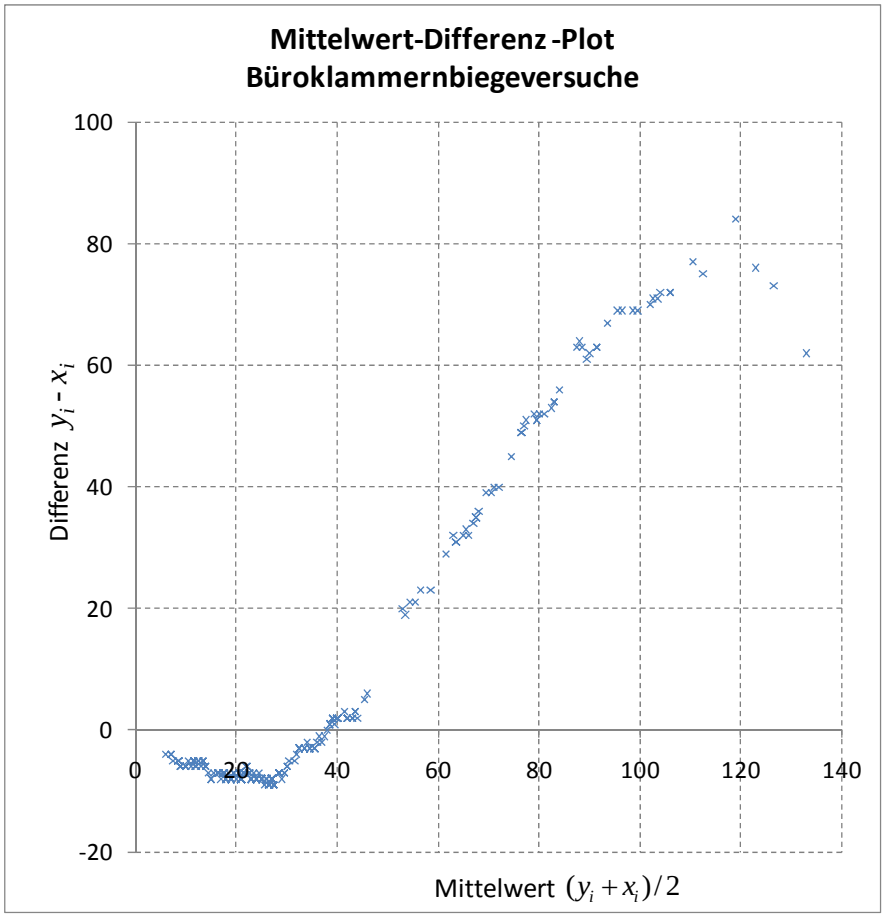
- Q-Q plots dienen zur Darstellung und dem Vergleich von zwei Datenreihen.
- Datenpunkte der beiden Datenreihen mit demselben Quantilwert werden aufgetragen.



Mittelwert-Differenz Plot

- Mittelwert-Differenz Plots dienen zur Darstellung und dem Vergleich von zwei Datenreihen.

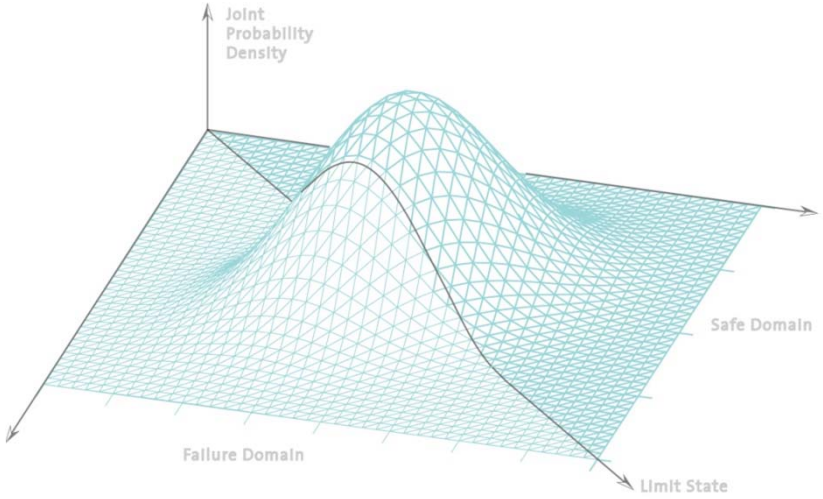
- Das Mittel $(y_i + x_i)/2$ wird über die Differenz $y_i - x_i$ aufgetragen.



y = grosse Klammern, x = kleine Klammern

Zusammenfassung Graphische Darstellung

Eindimensionales Streudiagramm	Veranschaulicht den Bereich und die Verteilung von Datenreihen entlang einer Achse, und zeigt Symmetrie.
Zweidimensionales Streudiagramm	Veranschaulicht den paarweisen Zusammenhang von Daten.
Histogramm	Stellt die Verteilung von Daten über einem Bereich von Datenreihen dar, zeigt Modalwert und Symmetrie.
Quantil-Plot	Stellt Median, Verteilung und Symmetrie dar.
Tukey Box Plot	Stellt Median, obere/untere Quartile, Symmetrie und Verteilung dar.
Q-Q Plot	Vergleicht zwei Datenreihen, relatives Bild.
Mittelwert-Differenz Plot	Vergleicht zwei Datenreihen, relatives Bild.



Statistik und Wahrscheinlichkeitsrechnung