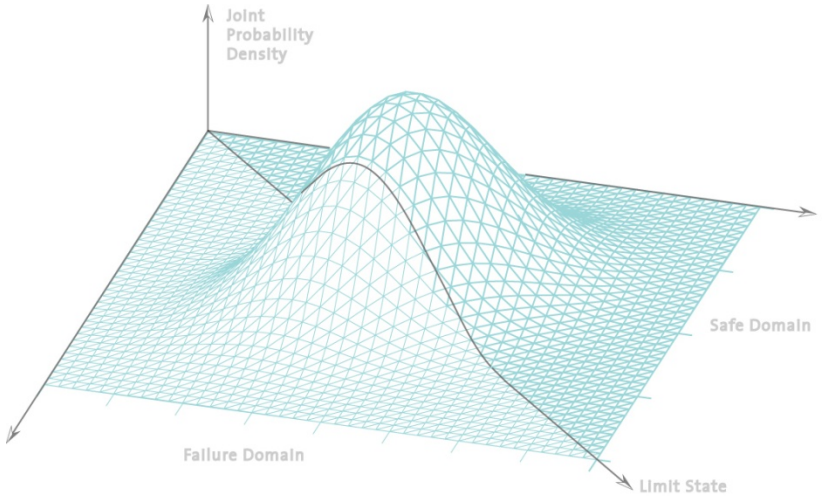
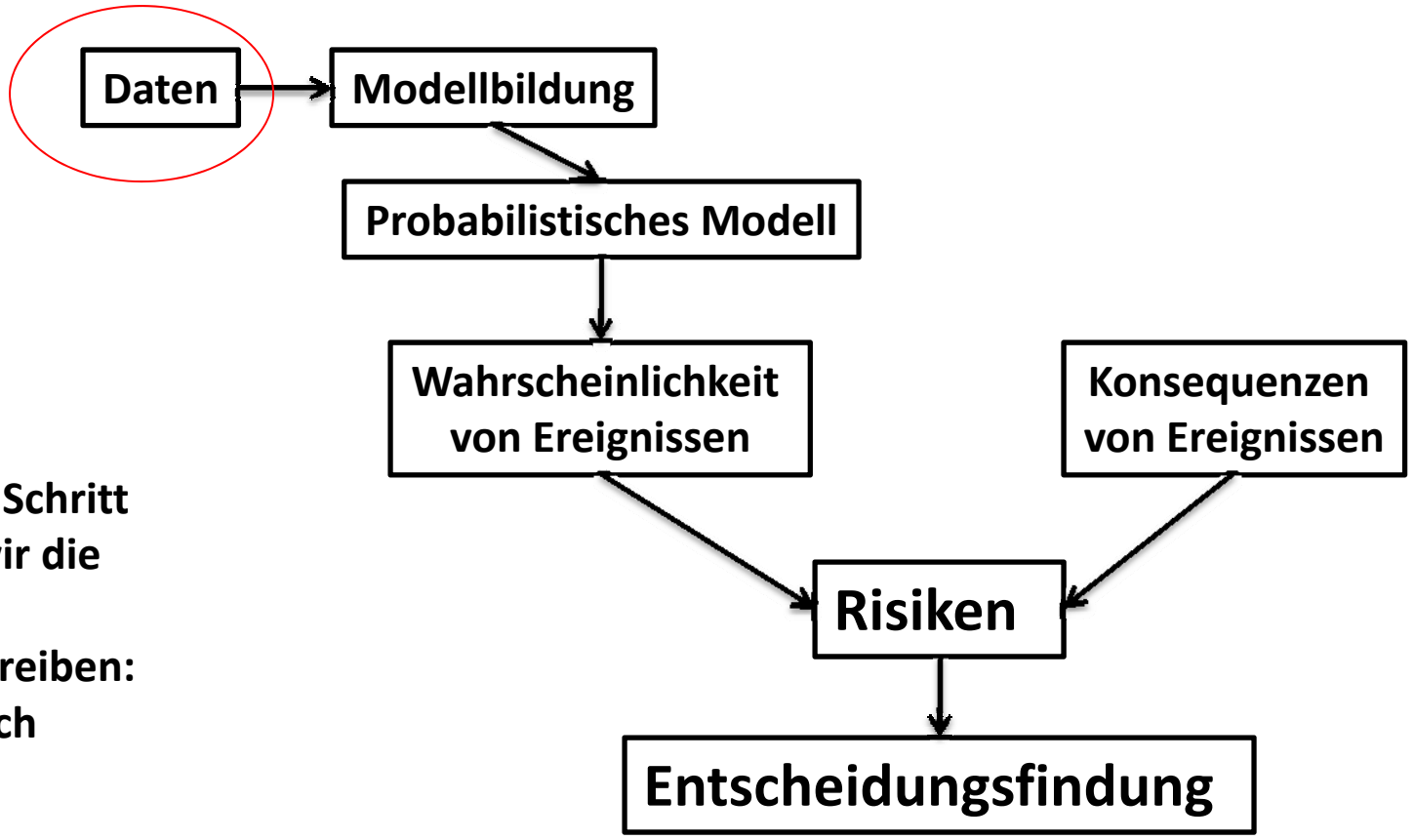


Statistik und Wahrscheinlichkeitsrechnung

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

170	191	184	184	182	176	183	164	178	183
190	176	175	170	171	180	176	177	187	170
190	171	183	182	178	180	185	175	180	184
175	169	183	176	179	172	176	180	183	182
173	165	175	190	160	189	174	184	191	165
170	165	178	180	176	185	187	174	187	184
183	166	177	189	197	174	166	186	184	174
178	183	180	176	185	178	185	185	184	183
190	186	183	183	178	188	185	181	175	171
175	170	168	178	185	184	187	162	170	183
175	174	187	176	184	183	184	195	180	178
183	187	160	200	170	179	160	179	180	
164	172	175	181	170	179	189	182	183	
176	164	175	176	188	185	190	179	175	
169	176	162	175	187	175	173	180	174	
178	180	175	185	182	182	168	183	170	
188	178	158	177	186	176	184	182	170	
187	191	158	173	158	183	178	165	174	
164	174	187	175	172	177	187	186	181	
183	178	172	183	176	173	187	175	175	

⇨ **Kennwerte**

⇨ **Grafiken**

Keine Annahmen – nur Beschreibung !!

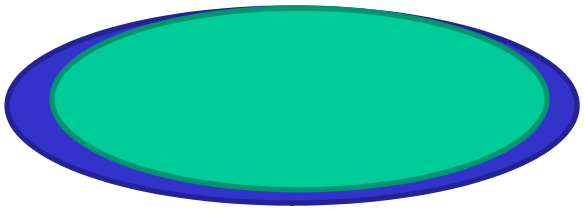
Vorbemerkung

- 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 = 258.

  Stichprobe von letzter Woche, n = 204.



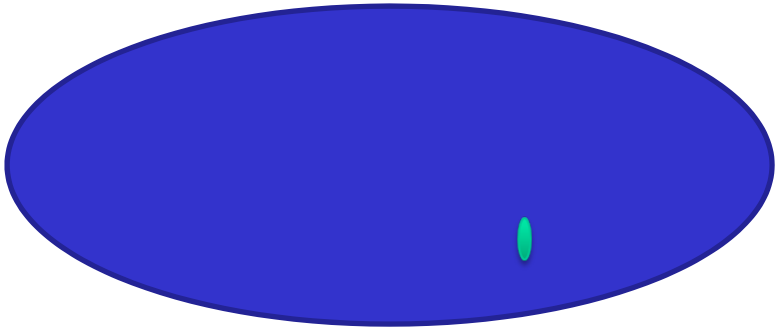
Vorbemerkung

- Stichprobe und Grundgesamtheit

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

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

  Stichprobe, $n = 204$



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

170	191	184	184	182	176	183	164	178	183
190	176	175	170	171	180	176	177	187	170
190	171	183	182	178	180	185	175	180	184
175	169	183	176	179	172	176	180	183	182
173	165	175	190	160	189	174	184	191	165
170	165	178	180	176	185	187	174	187	184
183	166	177	189	197	174	166	186	184	174
178	183	180	176	185	178	185	185	184	183
190	186	183	183	178	188	185	181	175	171
175	170	168	178	185	184	187	162	170	183
175	174	187	176	184	183	184	195	180	178
183	187	160	200	170	179	160	179	180	
164	172	175	181	170	179	189	182	183	
176	164	175	176	188	185	190	179	175	
169	176	162	175	187	175	173	180	174	
178	180	175	185	182	182	168	183	170	
188	178	158	177	186	176	184	182	170	
187	191	158	173	158	183	178	165	174	
164	174	187	175	172	177	187	186	181	
183	178	172	183	176	173	187	175	175	

⇨ **Kennwerte**

⇨ **Grafiken**

Keine Annahmen – nur Beschreibung !!

Datenbeschreibung

- Zusammenfassen zu nur einem Kennwert

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

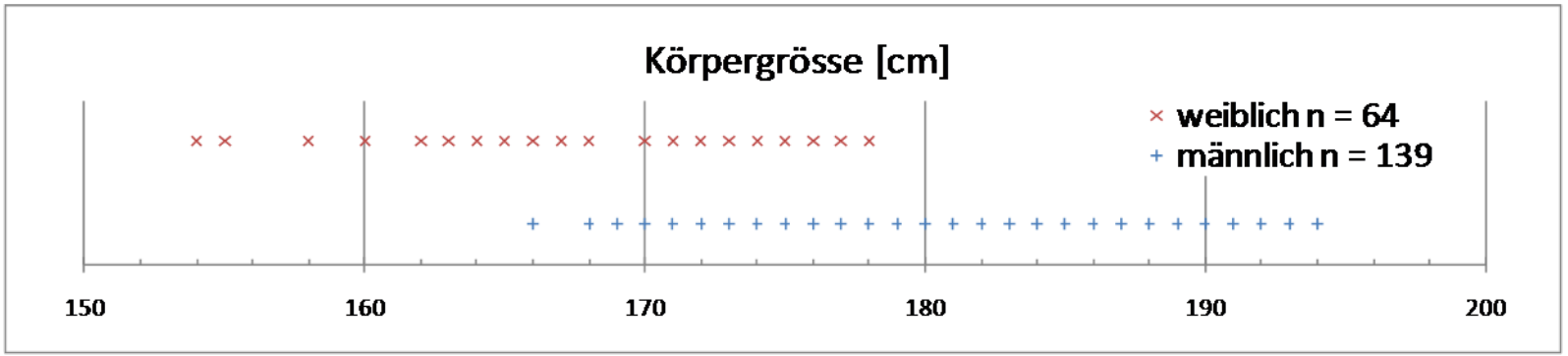
Für einen Datensatz: $$\mathbf{x} = \left(x_1, x_2, \ldots, x_n\right)^T$$

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

Datenbeschreibung

- Einfache grafische Darstellung von Stichproben

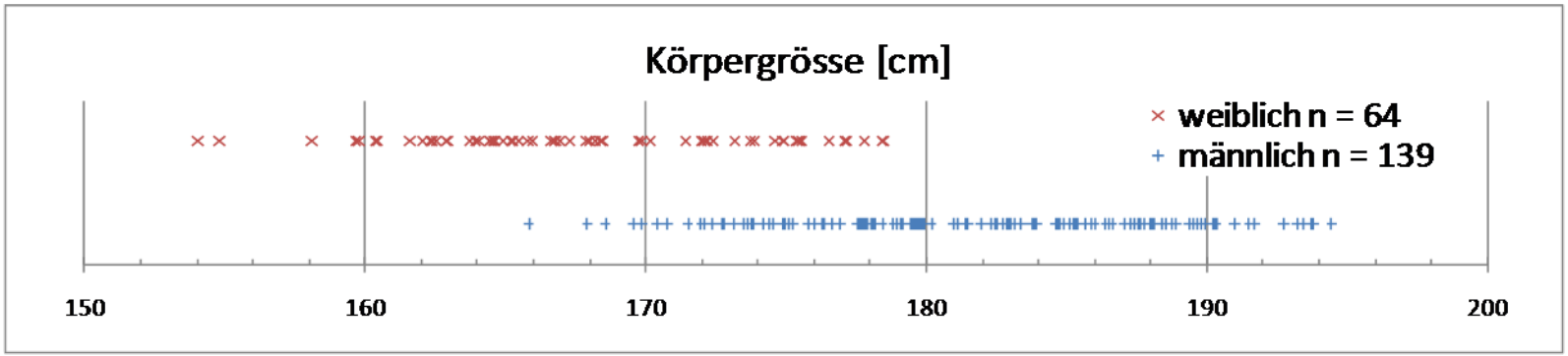
Eindimensionales Streudiagramm:



Datenbeschreibung

- Einfache grafische Darstellung von Stichproben

Eindimensionales Streudiagramm:



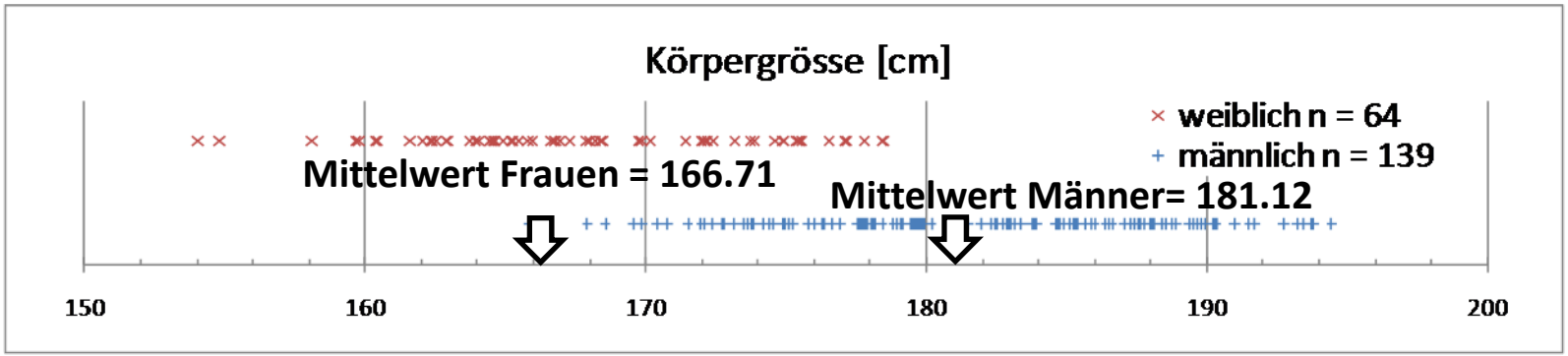
Guter Datenüberblick (Maximum, Minimum).

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

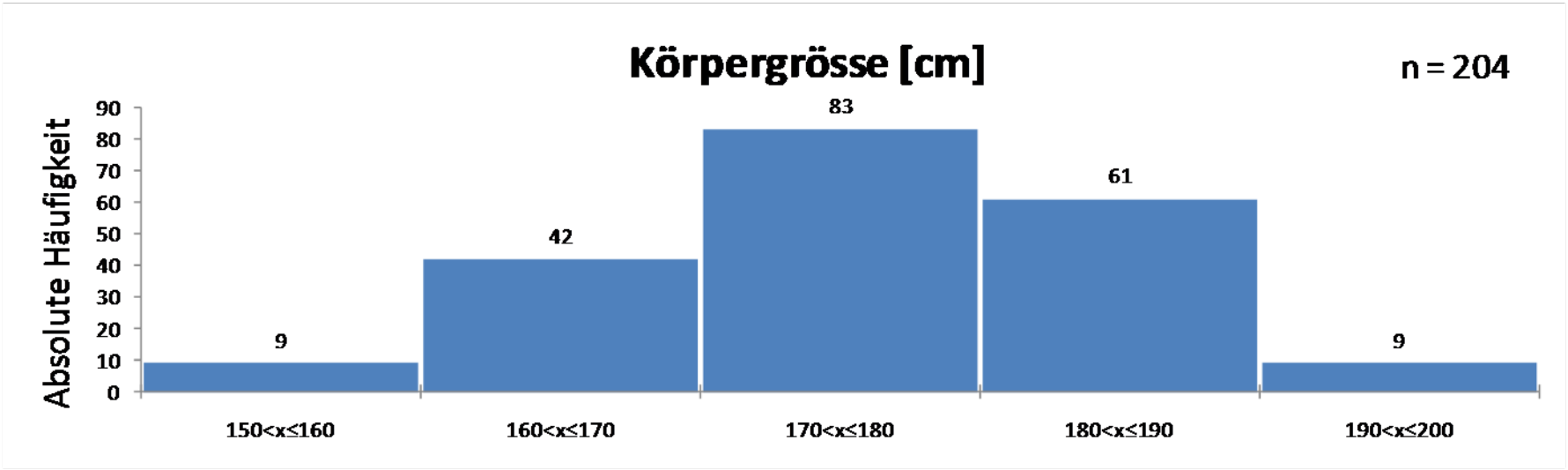
Intervall	Anzahl
150<x≤160	9
160<x≤170	42
170<x≤180	83
180<x≤190	61
190<x≤200	9
n =	204

Datenbeschreibung

Intervall	Anzahl
150<x≤160	9
160<x≤170	42
170<x≤180	83
180<x≤190	61
190<x≤200	9
n =	204

- Einfache grafische Darstellung von Stichproben

  Histogramm:

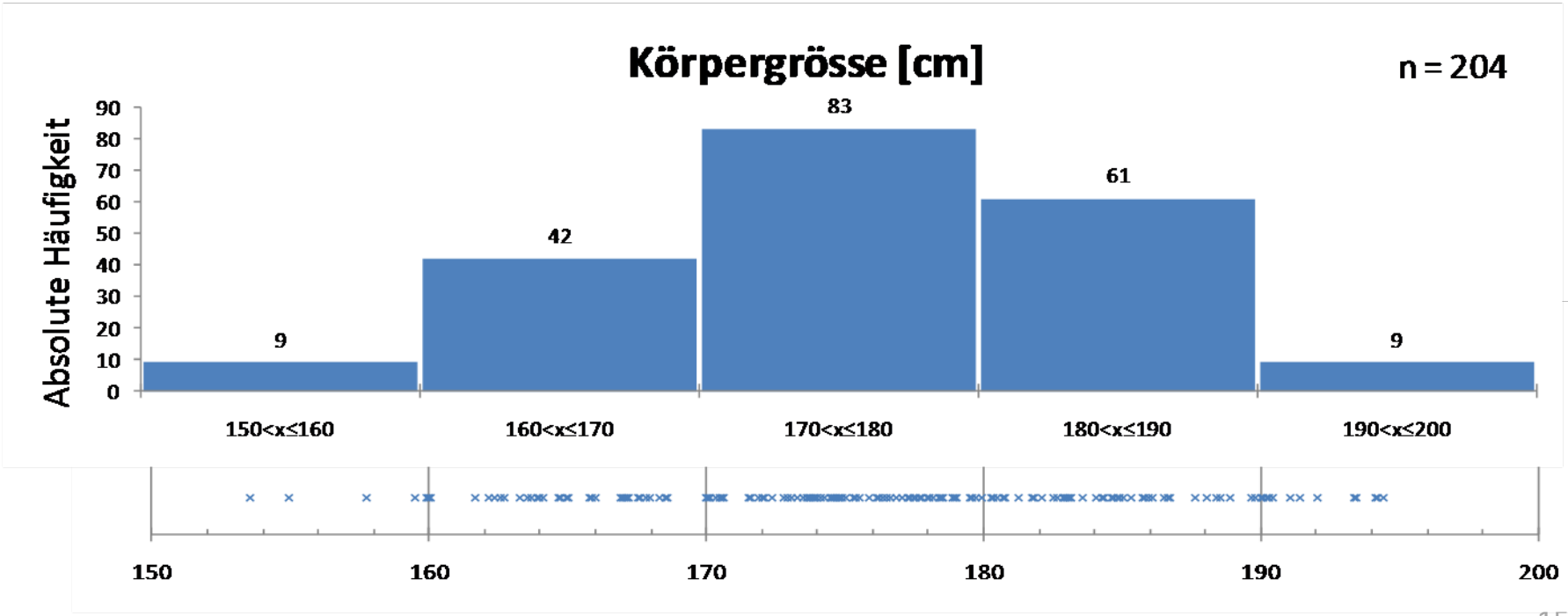


Datenbeschreibung

Intervall	Anzahl
150<x≤160	9
160<x≤170	42
170<x≤180	83
180<x≤190	61
190<x≤200	9
n =	204

- 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 $x_1^o \leq x_2^o \leq \ldots \leq x_n^o$.

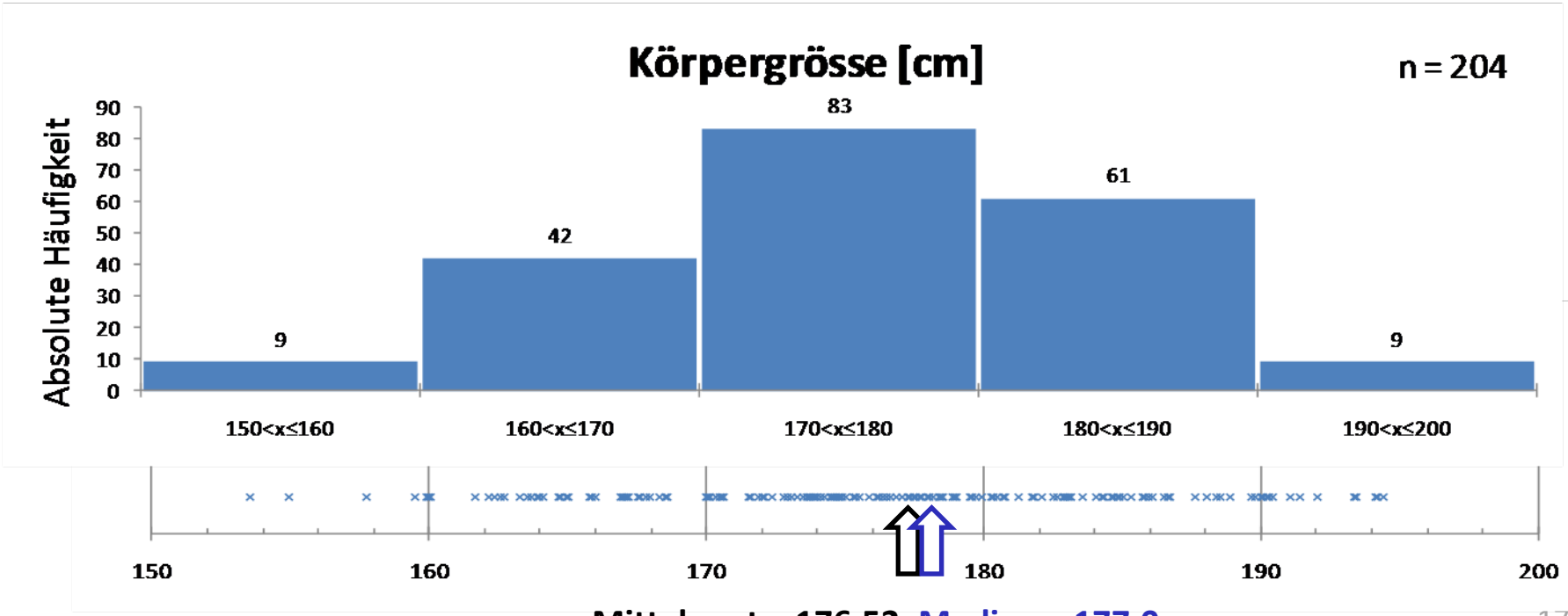
$$\tilde{x} = \begin{cases} x_{\frac{n+1}{2}} & \text{n ungerade} \\ \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1} \right) & \text{n gerade} \end{cases}$$

  - Beispiele: [23 30 31 33 120]

    [23 30 31 33]

Datenbeschreibung

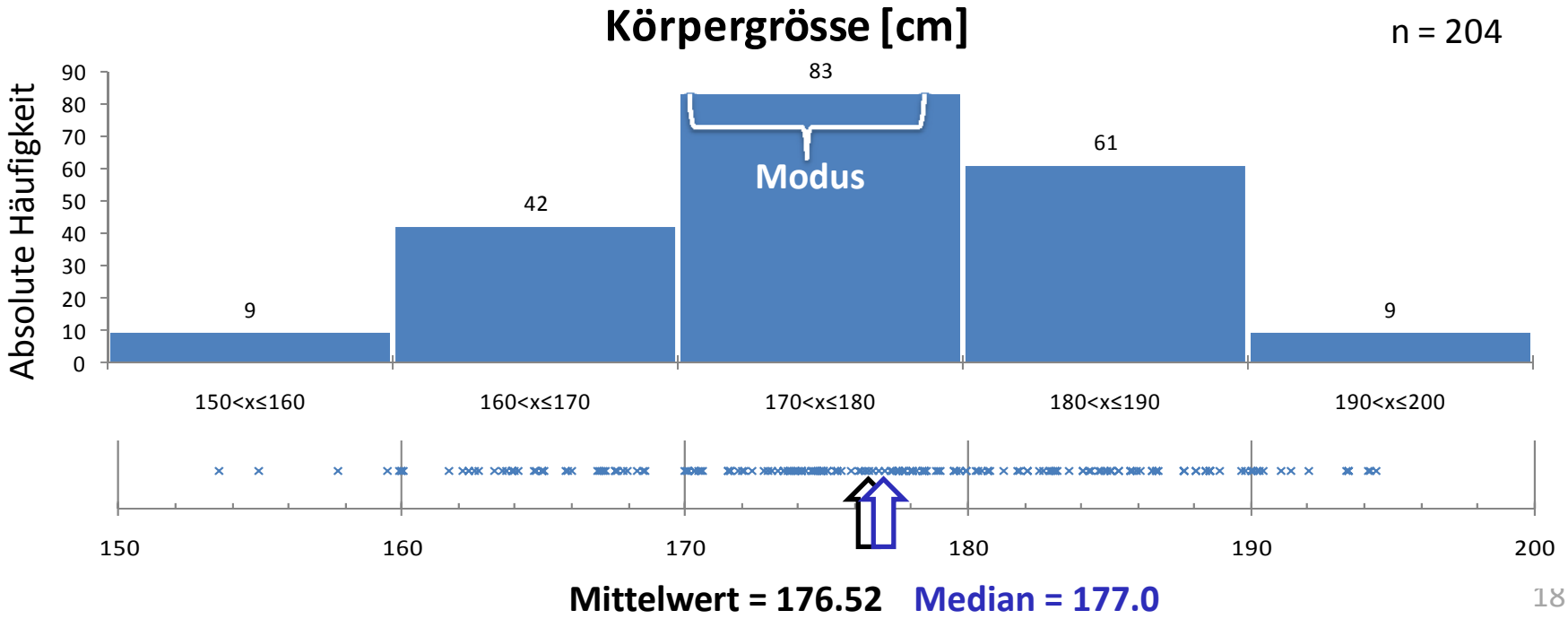
- Neben dem Mittelwert gibt es noch andere sog. Lageparameter:
  - Der **Median** oder Zentralwert $\tilde{x}$ der Stichprobe ist der mittlere Wert einer nach der Grösse geordneten Stichprobe $x_1^o \leq x_2^o \leq \ldots \leq x_n^o$.



Mittelwert = 176.52 **Median = 177.0**

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.



Datenbeschreibung

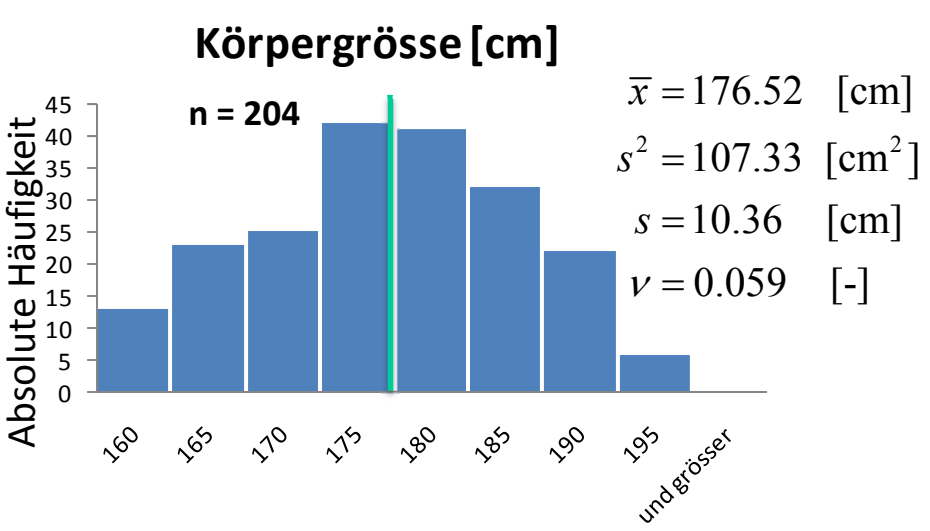
- Streuungsparameter – Streuung um den Mittelwert
  - Die Varianz der Stichprobe $$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$
  - Die Standardabweichung der Stichprobe $$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$
  - Der Variationskoeffizient der Stichprobe (relative Streuung, COV)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

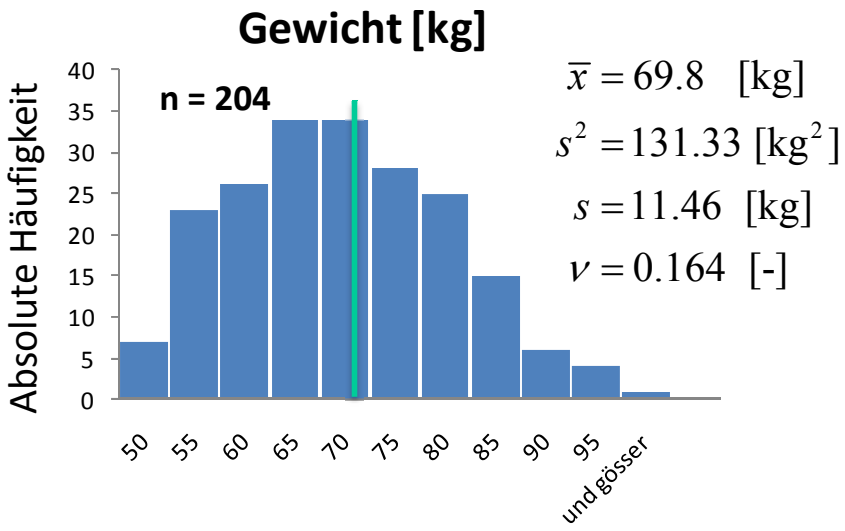
Datenbeschreibung

- 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





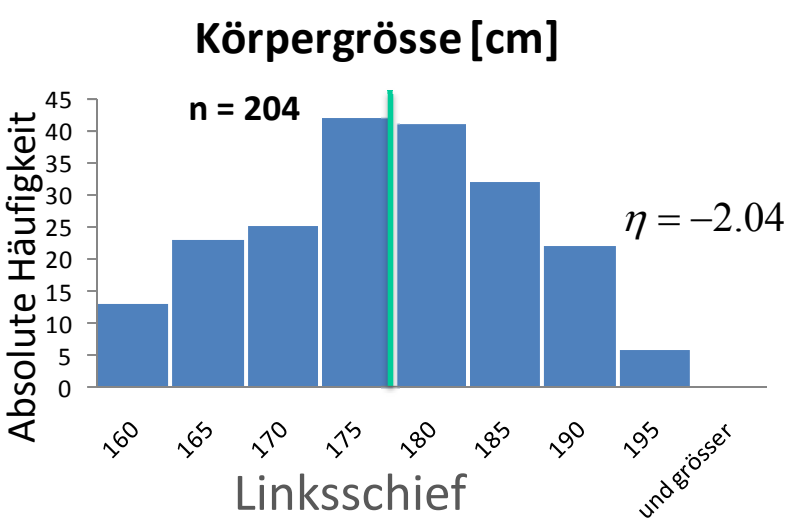
Datenbeschreibung

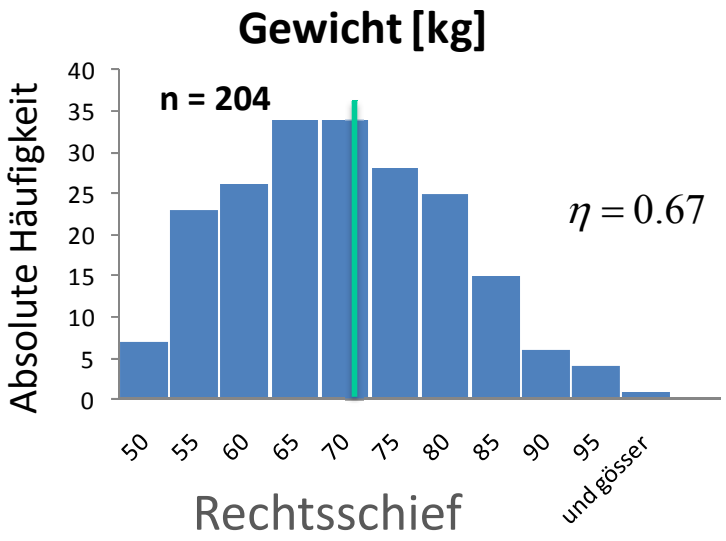
- Streuungsparameter – Streuung um den Mittelwert

  - 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

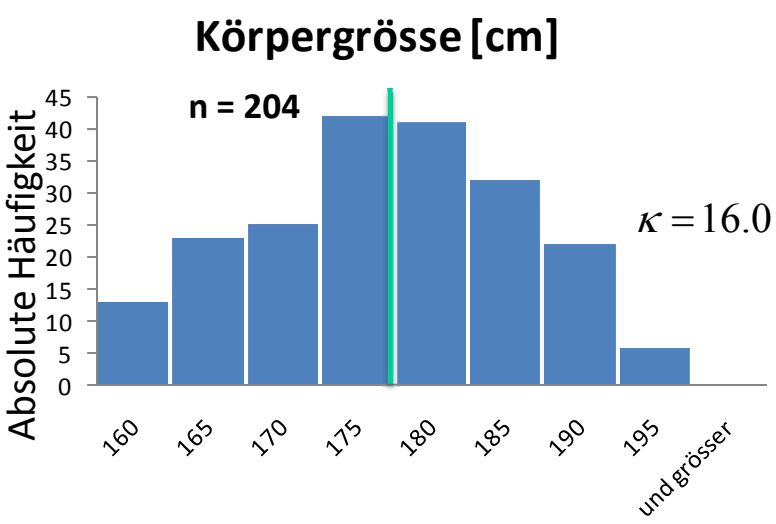
- Streuungsparameter – Streuung um den Mittelwert

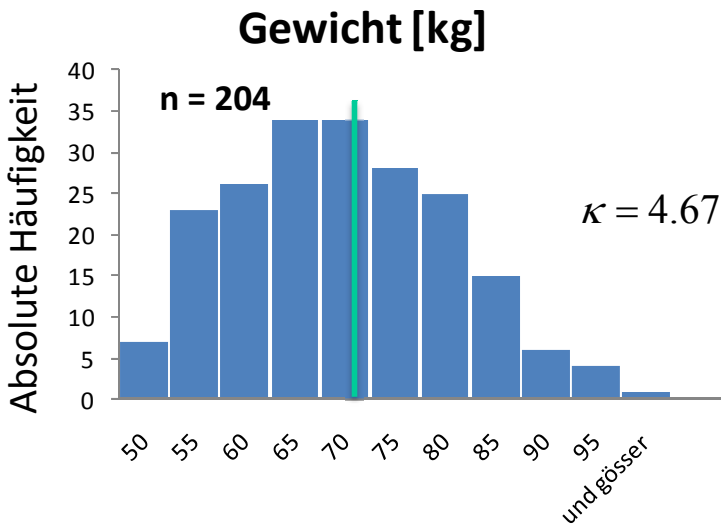
  - Kurtosis der Stichprobe:

    -> Mass für die Spitzigkeit / Gipfligkeit

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

Beispiel

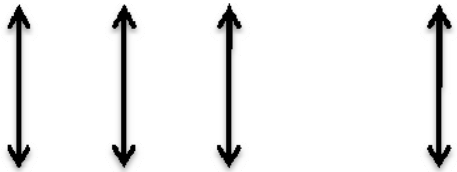




Datenbeschreibung

- Beschreibung von paarweise beobachteten Eigenschaften

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

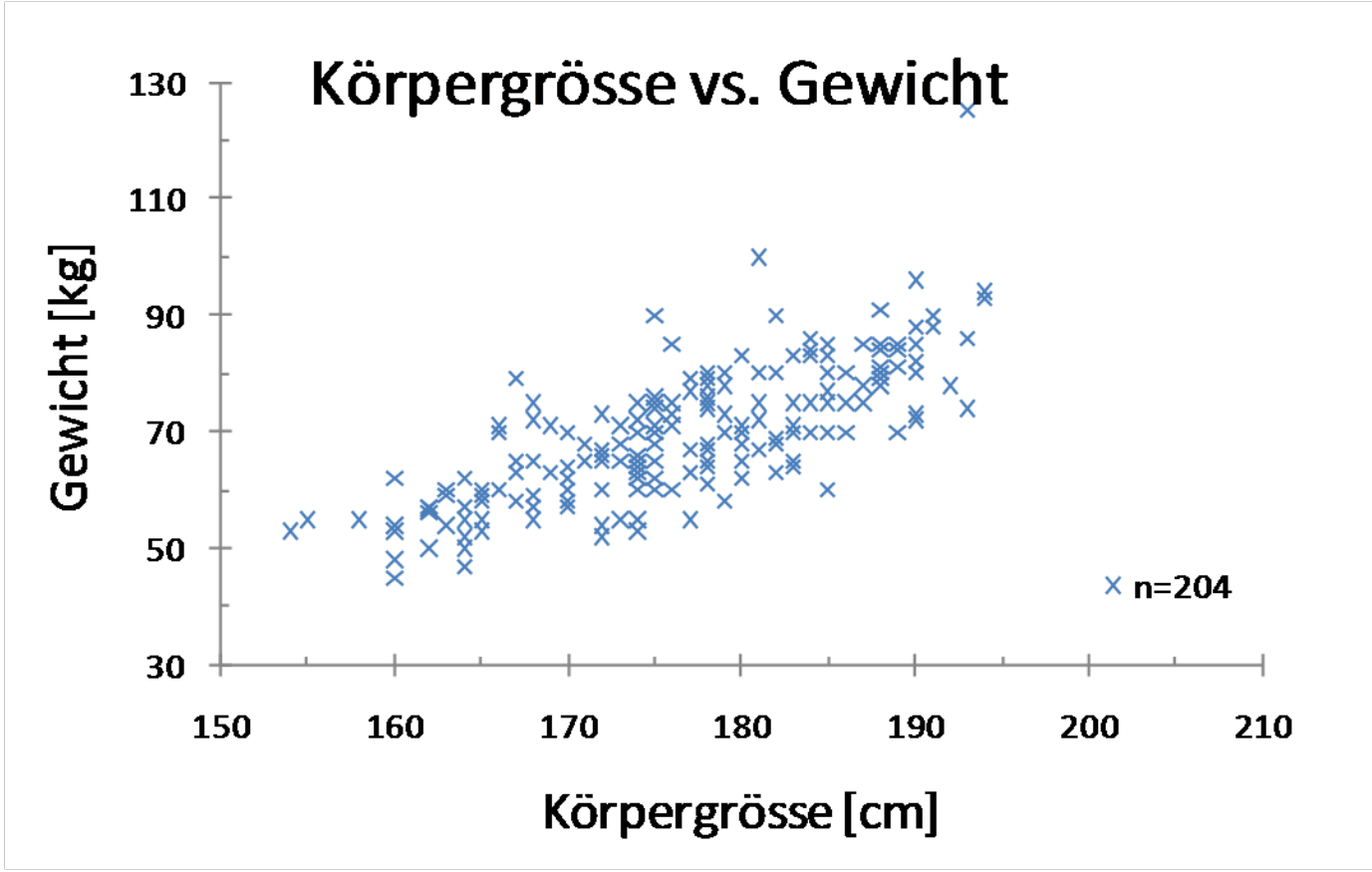


$$\mathbf{y} = (y_1, y_2, y_3, \ldots, y_n)^T$$

Datenbeschreibung

- Beschreibung von paarweise beobachteten Eigenschaften

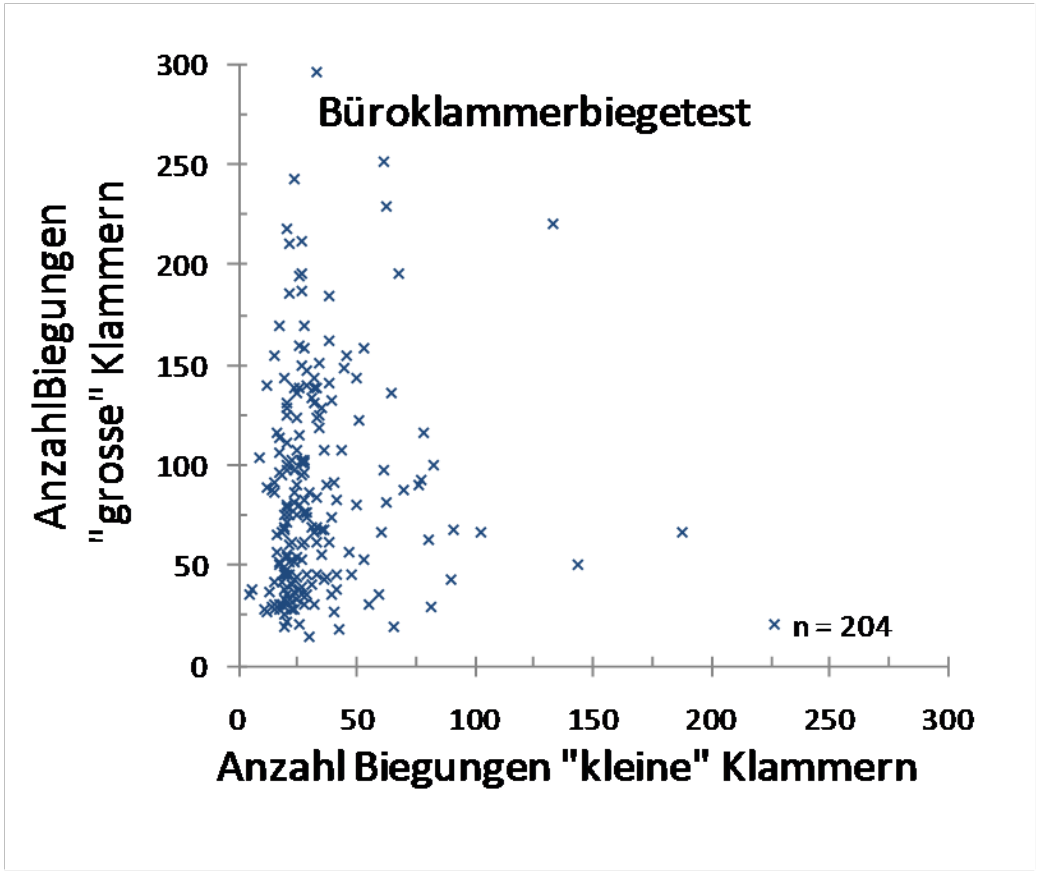
  Das zweidimensionale Streudiagramm



Datenbeschreibung

- Beschreibung von paarweise beobachteten Eigenschaften

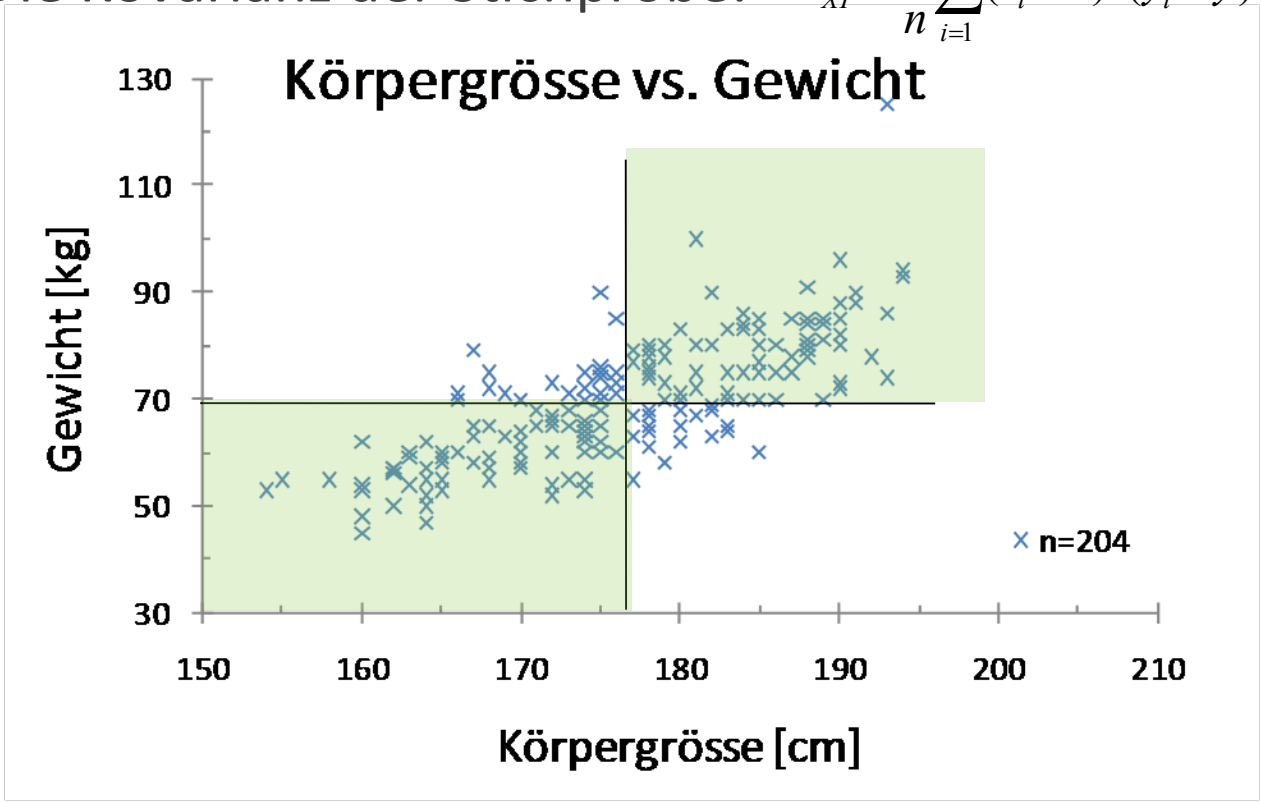
  Das zweidimensionale Streudiagramm



Datenbeschreibung

- Beschreibung von paarweise beobachteten Eigenschaften

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



$x \mathrel{\hat{=}}$ Körpergrösse

$\bar{x} = 176.52$ cm

$y \mathrel{\hat{=}}$ Gewicht

$\bar{y} = 69.80$ kg

Datenbeschreibung

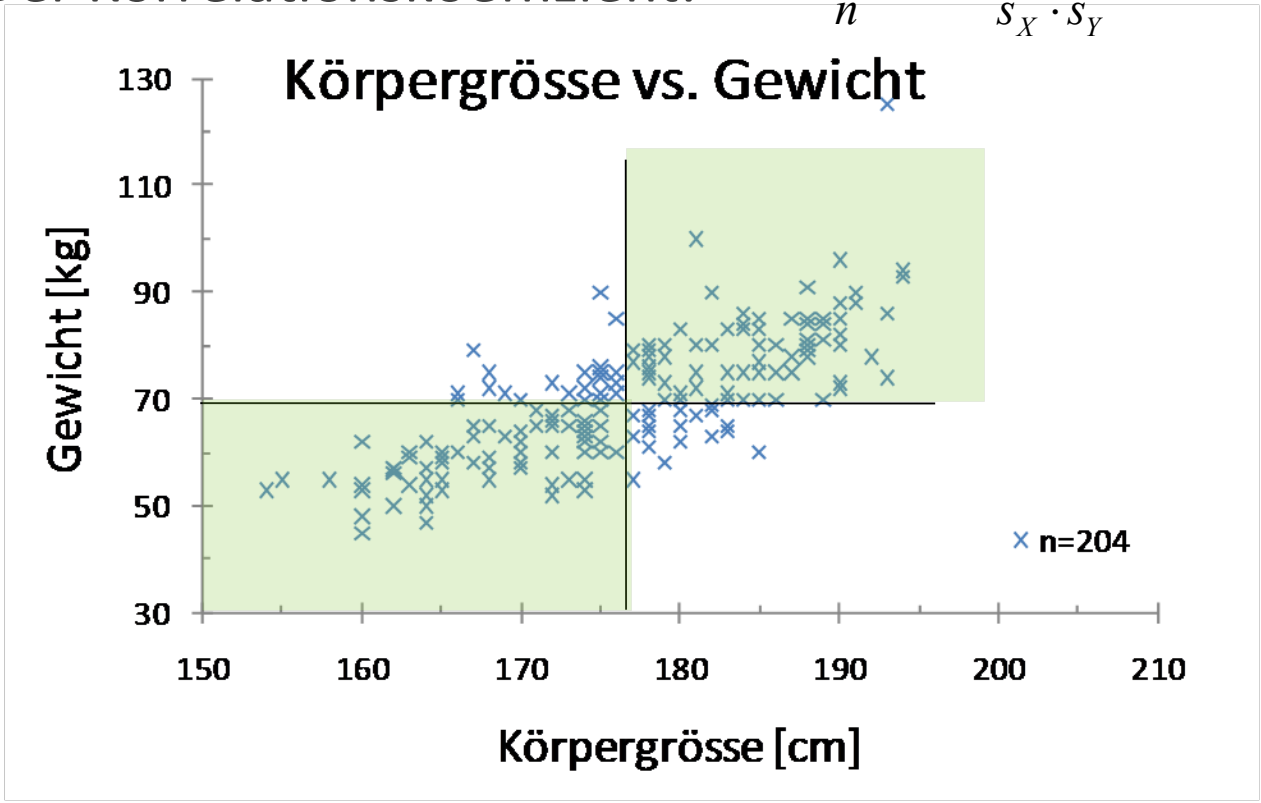
- Beschreibung von paarweise beobachteten Eigenschaften
- Die Kovarianz der Stichprobe: $s_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})$
- Der Korrelationskoeffizient der Stichprobe:

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

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

Datenbeschreibung

- Beschreibung von paarweise beobachteten Eigenschaften
- Der Korrelationskoeffizient: $r_{XY} = \frac{1}{n} \frac{\sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{s_X \cdot s_Y} = 0.693$



$x \mathrel{\hat{=}}$ Körpergrösse

$\bar{x} = 176.52$ cm

$y \mathrel{\hat{=}}$ Gewicht

$\bar{y} = 69.80$ 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

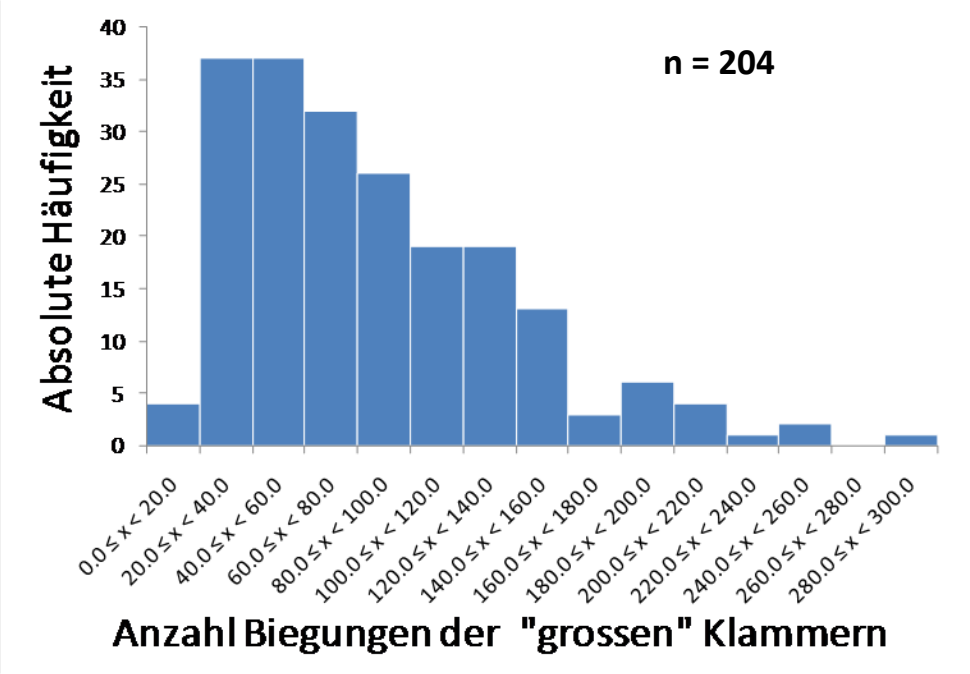
- Beispiel: Ihre Büroklammerdaten vom letzten Mal
  „grosse“ Klammern, Stichprobenumfang n = 204,
  Maximalwert 301, Minimalwert 9.

  Einteilung in 15 Intervalle; [0,20); [20,40); [40,60);… ; [300,320)

Histogramm

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

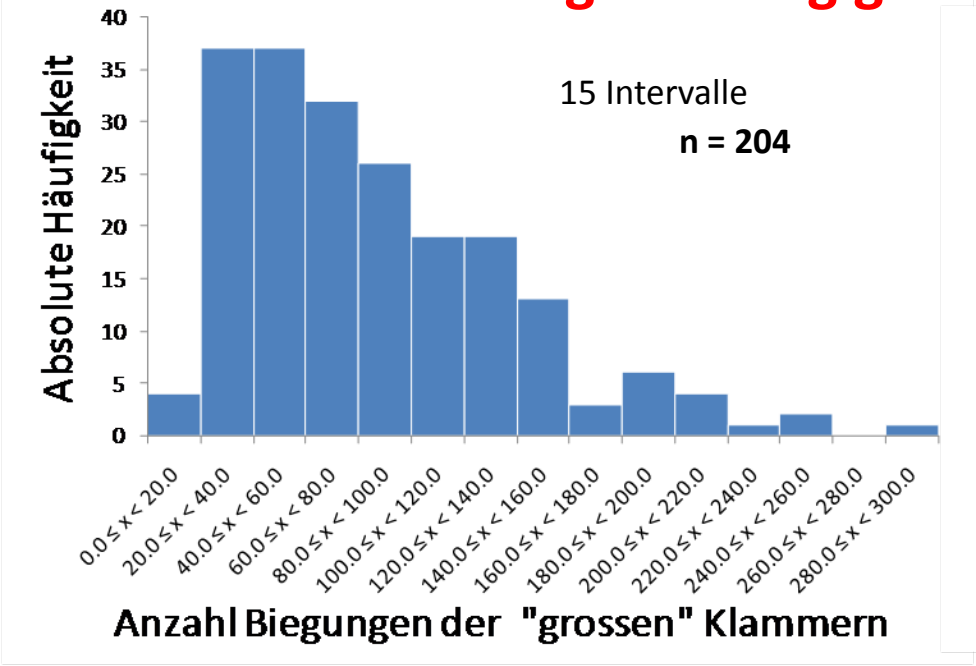
- Beispiel:

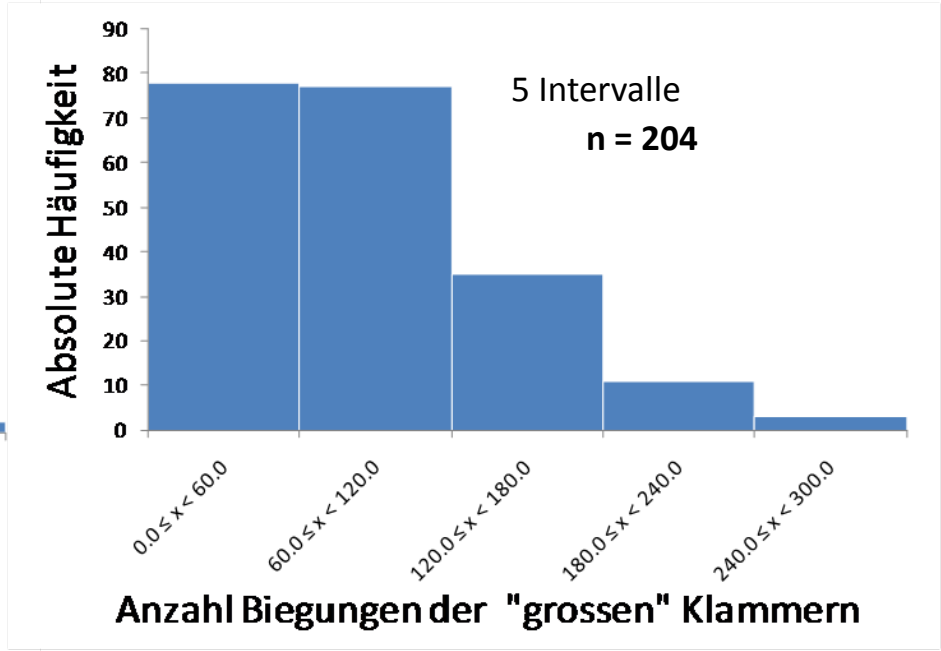


Histogramm

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

Aussage abhängig von der Anzahl der Intervalle!





Histogramm

- Prinzip:
  - Aufteilung der Stichprobe in *k* Grössen-Intervalle
  - Auftragen der Häufigkeit je Intervall
  - **Faustregel für die Anzahl der Intervalle:** $k = 1 + 3.3 \log(n)$

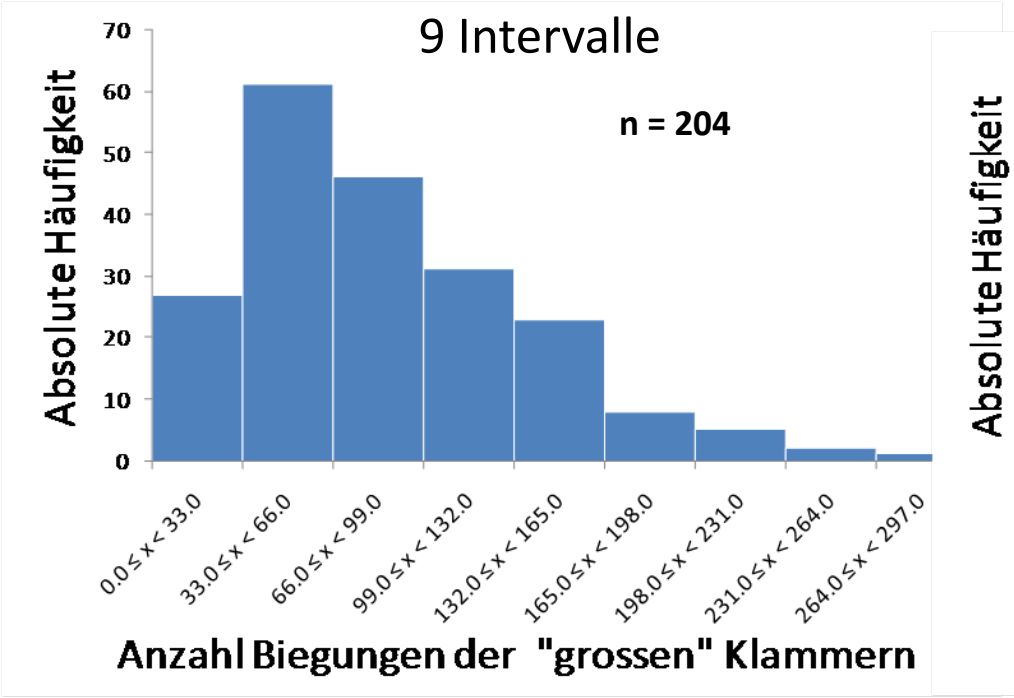
- Beispiel: Büroklammerdaten „grosse“ Klammern,
  Stichprobenumfang n = 204, Wertebereich [15,296]

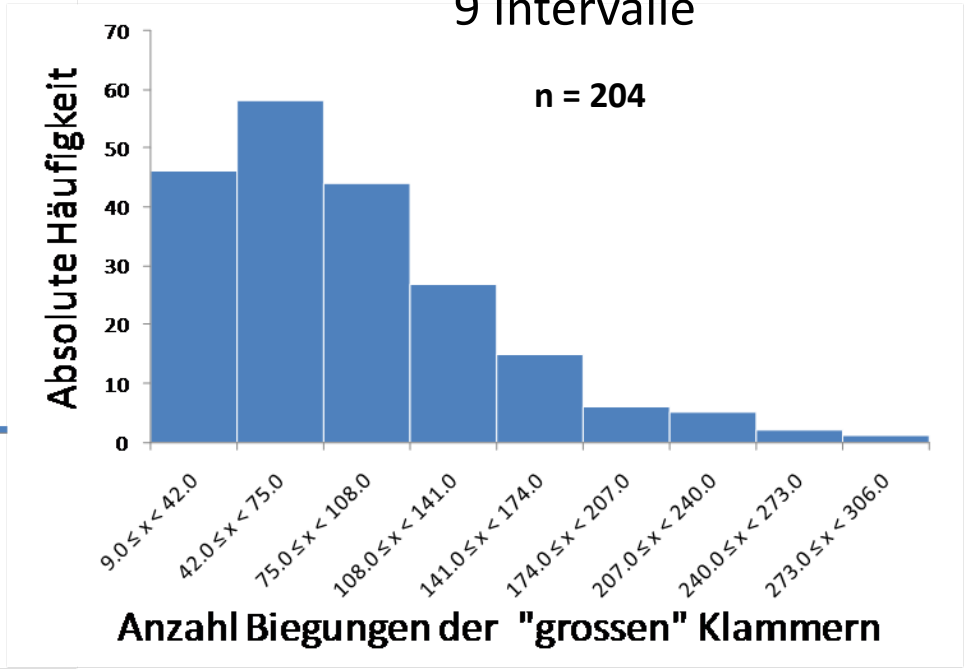
$$k = 1 + 3.3 \log(204) = 8.62 \cong 9 \text{ Intervalle}$$

[0,33); [33,66); [66,99);… ; [297,330)

oder [9,42); [42,75); [75,108);… ; [306,339) ?

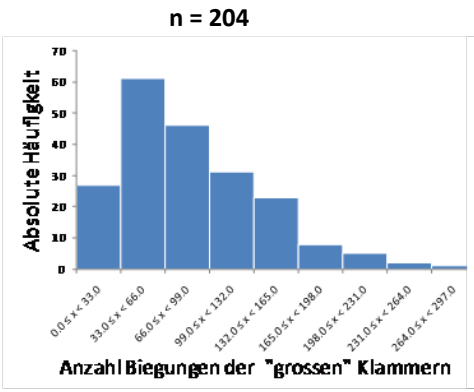
Histogramm

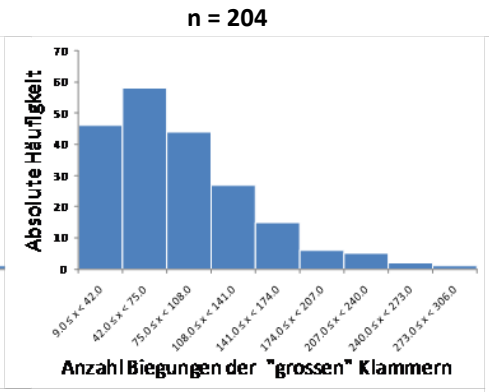

9 Intervalle
n = 204
Absolute Häufigkeit
70
60
50
40
30
20
10
0
0.0 ≤ x < 33.0
33.0 ≤ x < 66.0
66.0 ≤ x < 99.0
99.0 ≤ x < 132.0
132.0 ≤ x < 165.0
165.0 ≤ x < 198.0
198.0 ≤ x < 231.0
231.0 ≤ x < 264.0
264.0 ≤ x < 297.0
Anzahl Biegungen der "grossen" Klammern

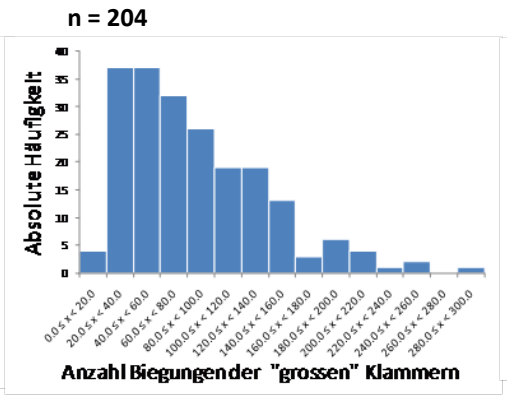

9 Intervalle
n = 204
Absolute Häufigkeit
70
60
50
40
30
20
10
0
9.0 ≤ x < 42.0
42.0 ≤ x < 75.0
75.0 ≤ x < 108.0
108.0 ≤ x < 141.0
141.0 ≤ x < 174.0
174.0 ≤ x < 207.0
207.0 ≤ x < 240.0
240.0 ≤ x < 273.0
273.0 ≤ x < 306.0
Anzahl Biegungen der "grossen" Klammern

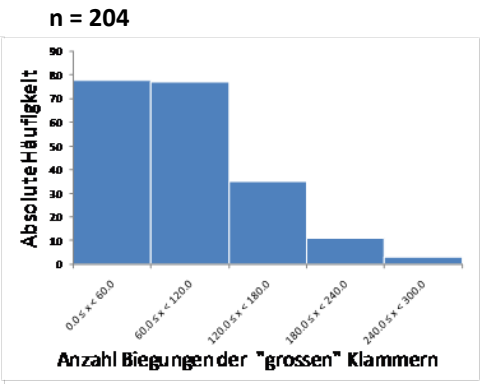
Histogramm

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



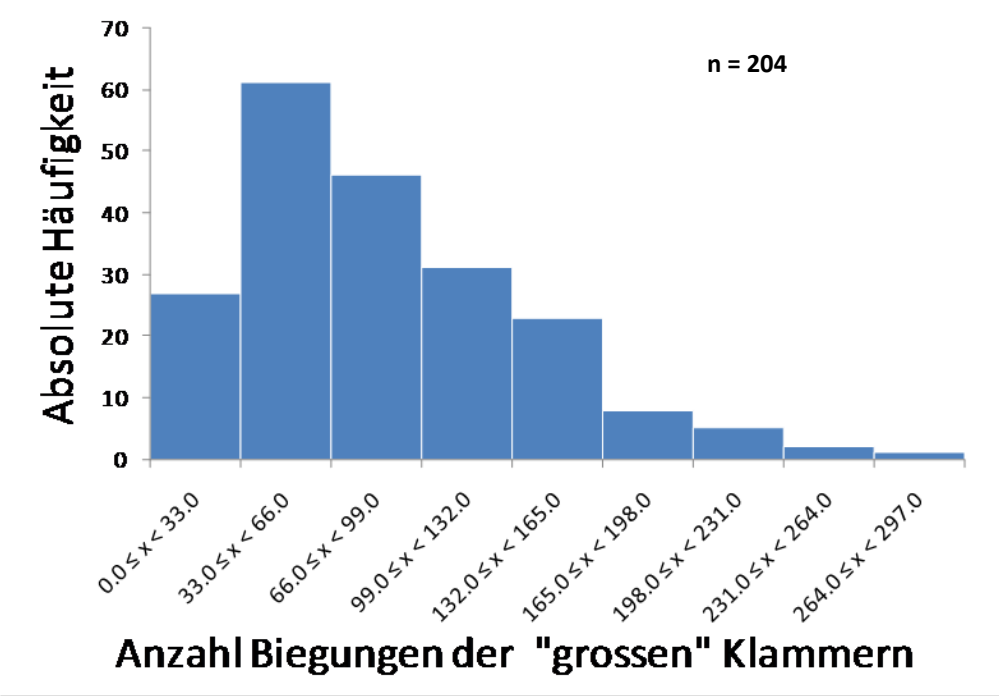






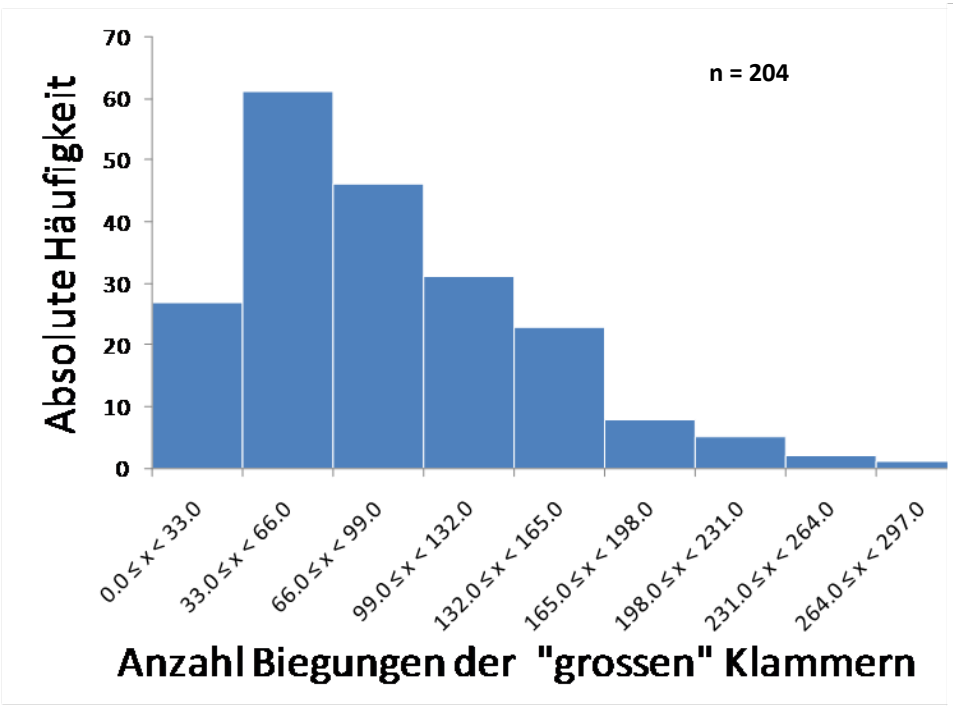
Histogramm

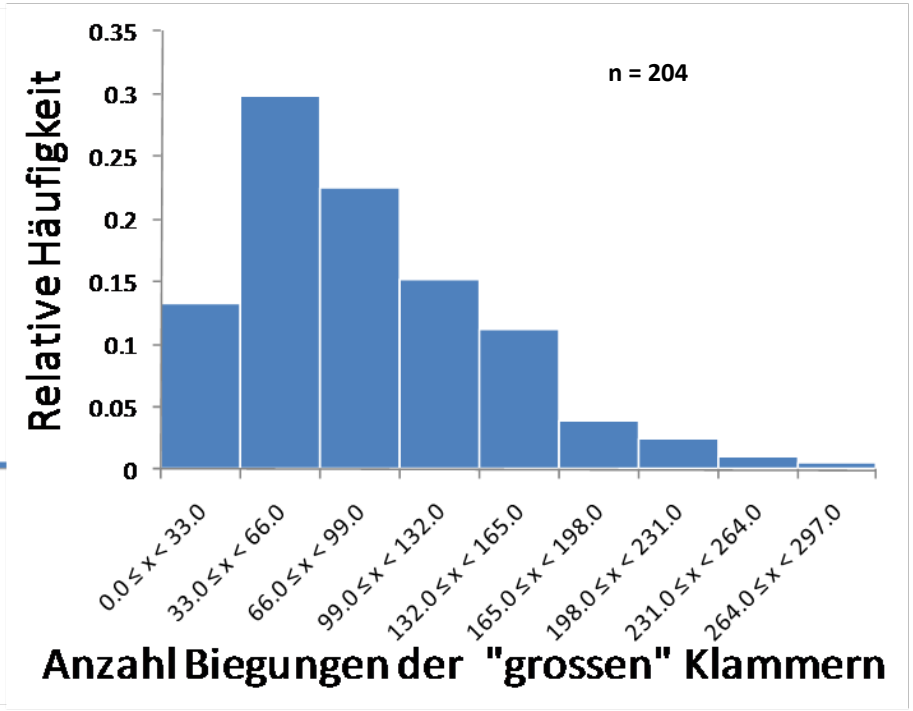
- Bisher haben wir die absolute Häufigkeit betrachtet.



Histogramm

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

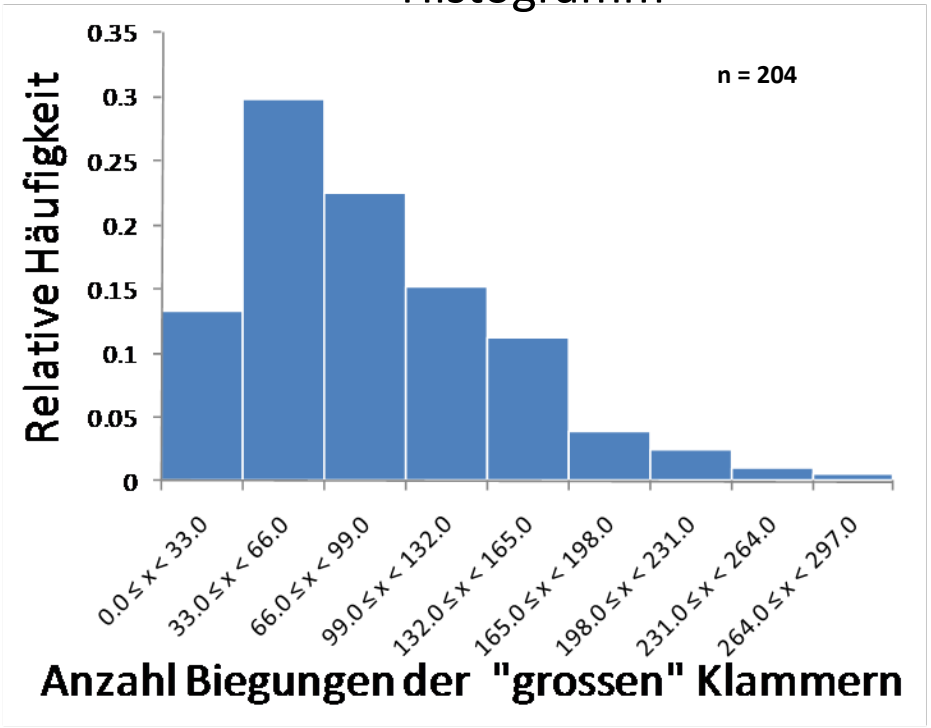




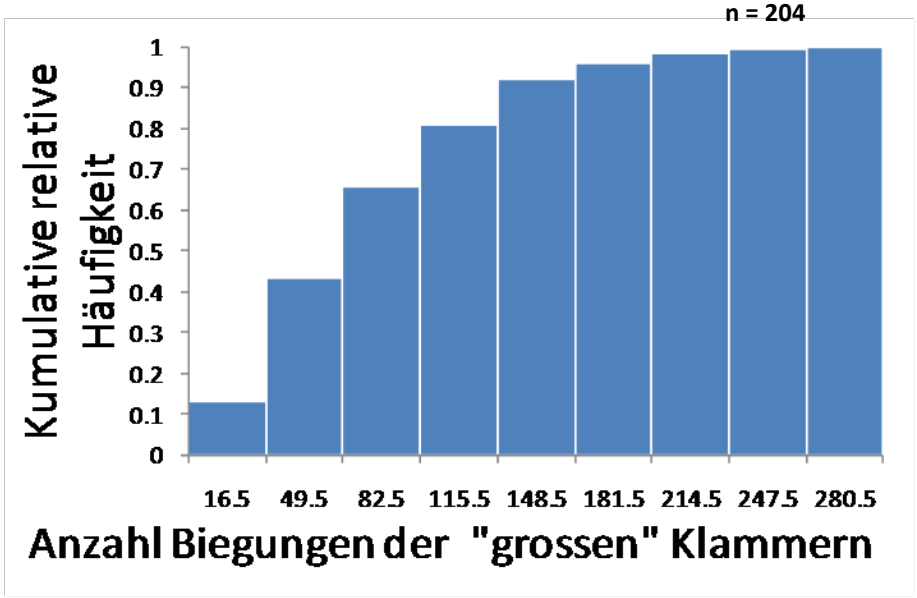
Histogramm

- Eine Spielart des Histogramms ist das kumulative Häufigkeitsdiagramm.

Histogramm

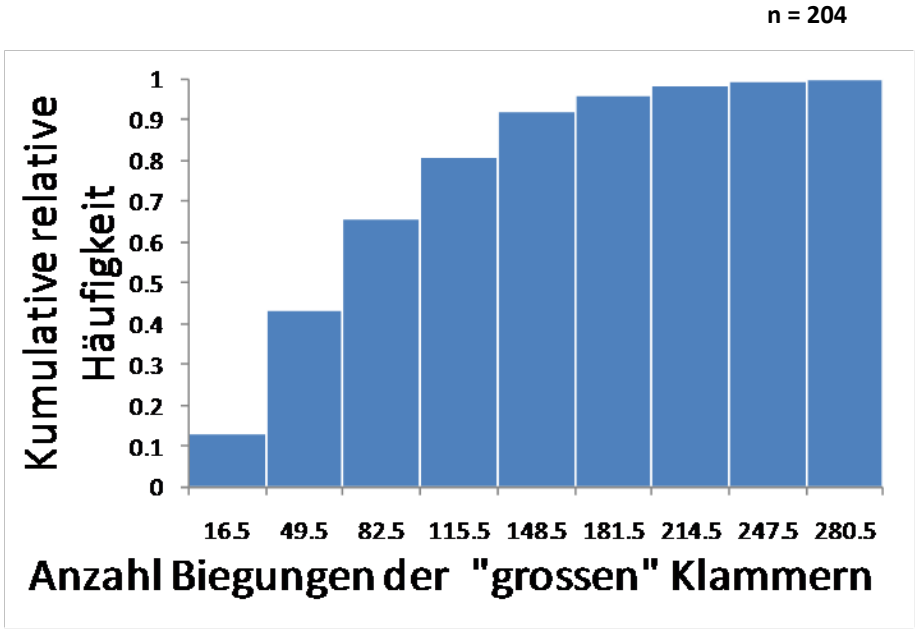


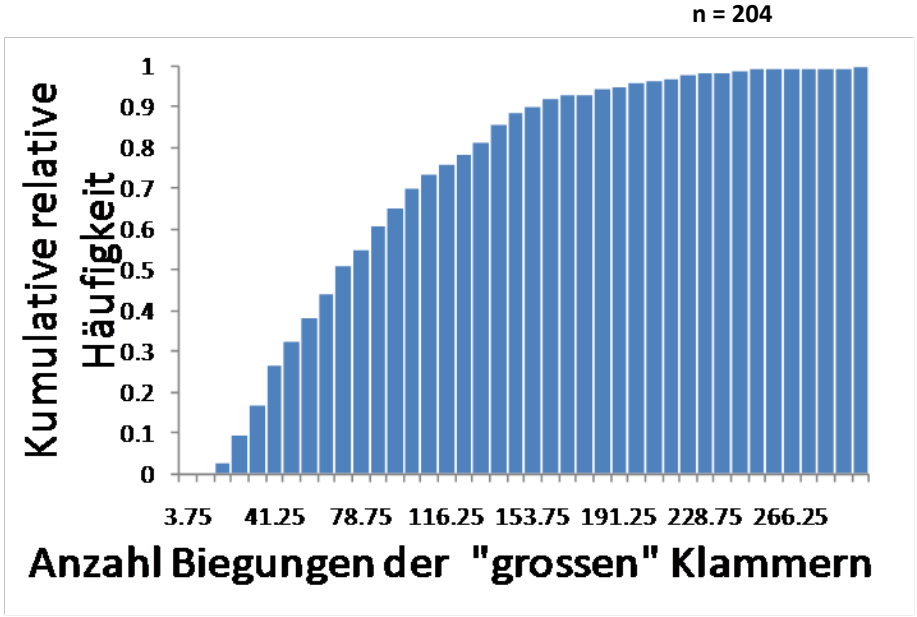
kumulatives Häufigkeitsdiagramm



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

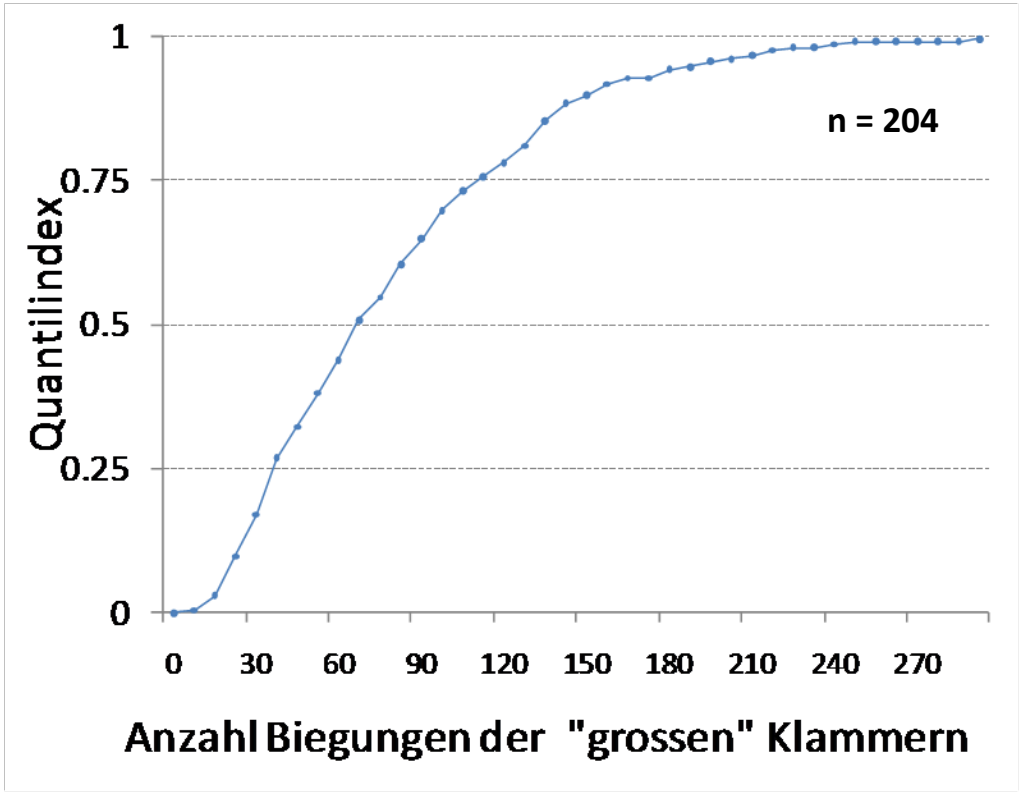
Das Quantil ist für eine gegebene Anzahl an Beobachtungen wie folgt definiert:

- 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: $x_1^o \leq x_2^o \leq \ldots \leq x_n^o$
- 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

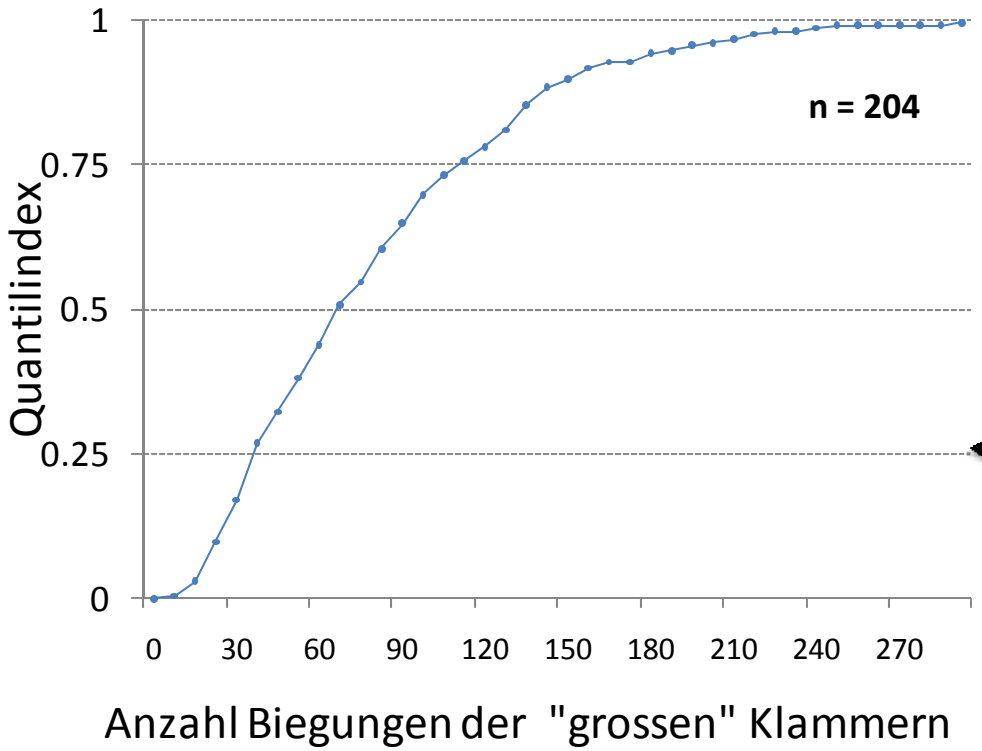
- Quantil-Plots werden durch Auftragen der Daten und der Quantilindizes gebildet.



i	$i/n+1$	x_i
1	0.0049261	6
2	0.0098522	8
3	0.0147783	9
4	0.0197044	10
5	0.0246305	10
6	0.0295567	10
7	0.0344828	11
8	0.0394089	12
9	0.0443350	12
.	.	.
.	.	.

Quantil - Plot

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

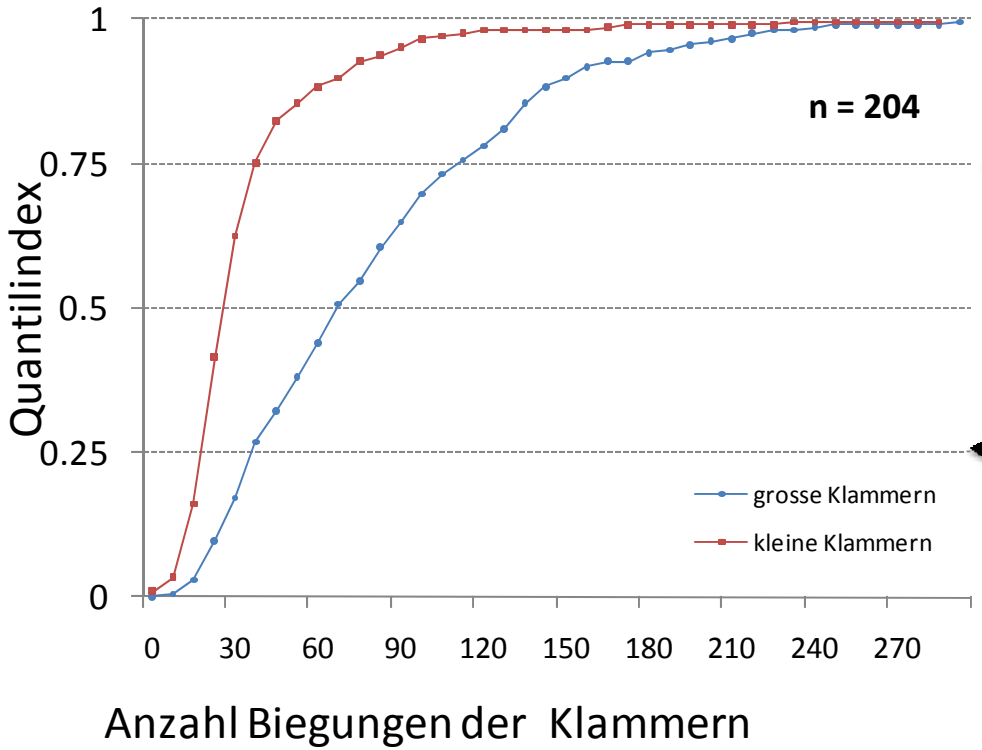


oberes Quartil = 0.75-Quantil

unteres Quartil = 0.25-Quantil

Quantil - Plot

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



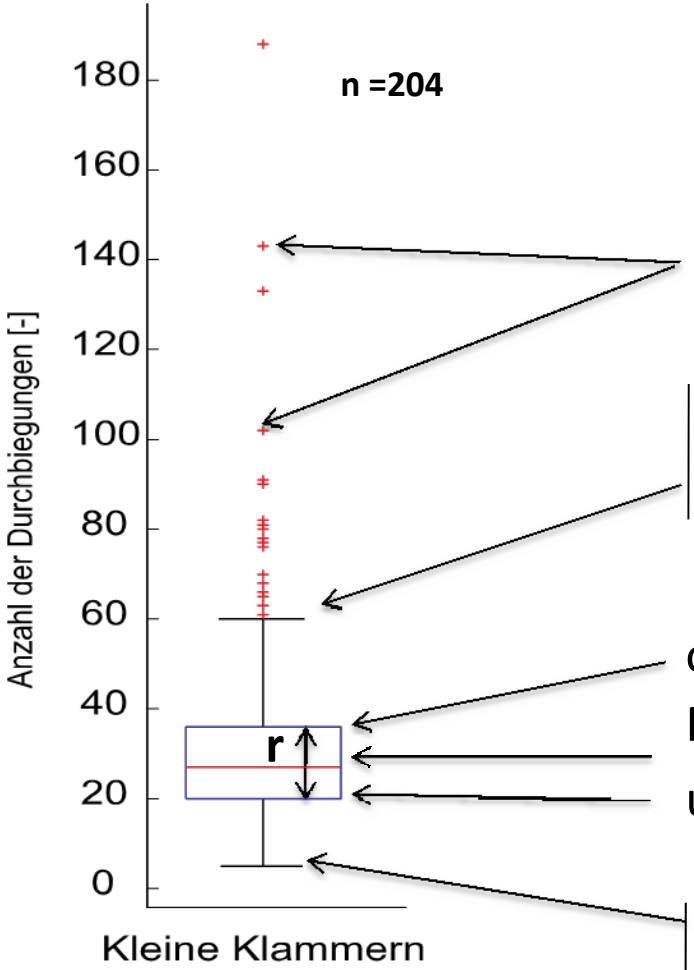
oberes Quartil = 0.75-Quantil

unteres Quartil = 0.25-Quantil

Tukey Box Plot

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

Tukey Box Plot



Ausreisser

oberer Nachbarschaftswert
grösste Beobachtung kleiner/gleich
oberes Quartil + 1.5 * r

r = interquartile Differenz

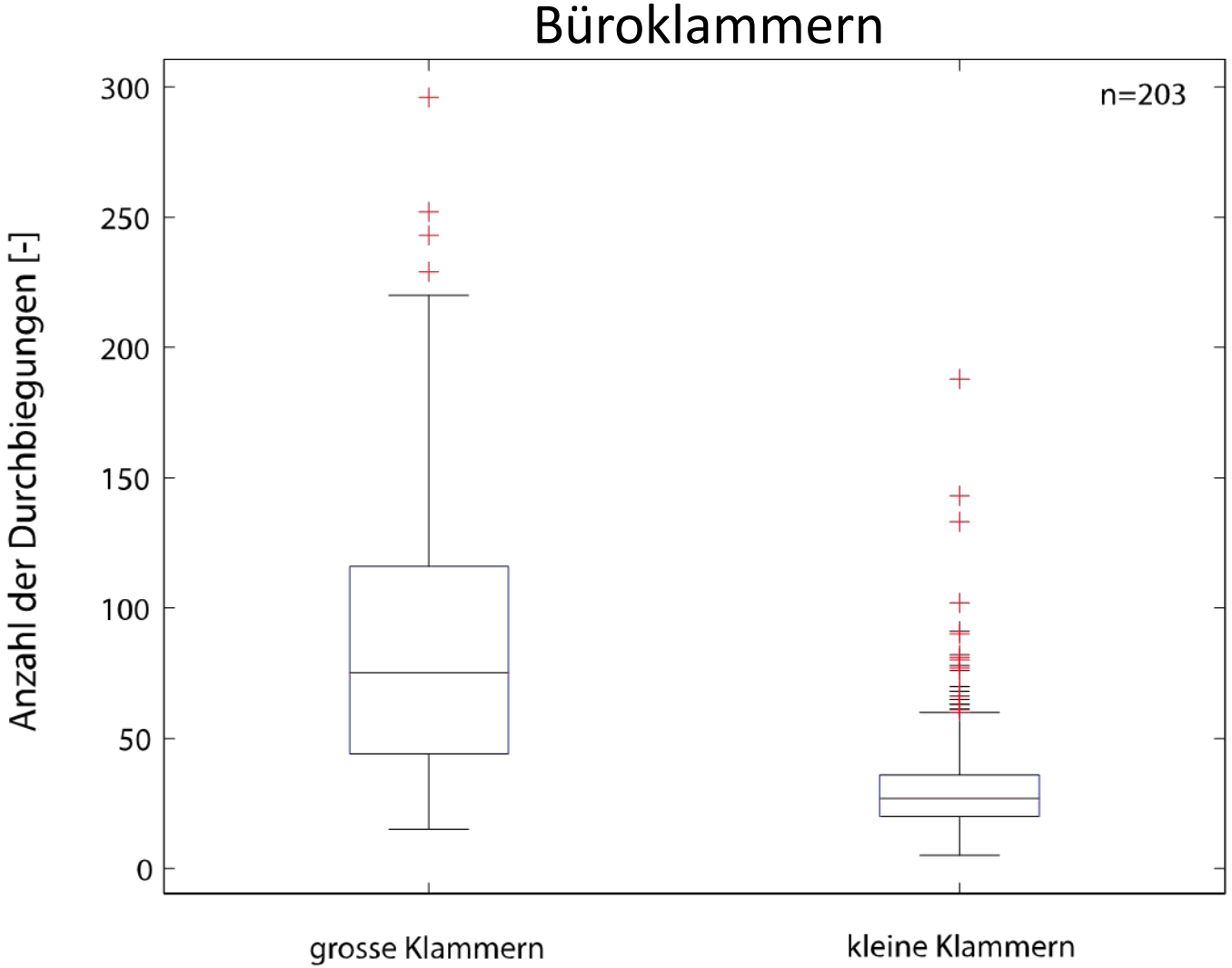
oberes Quartil = 0.75-Quantil

Median = 0.50-Quantil

unteres Quartil = 0.25-Quantil

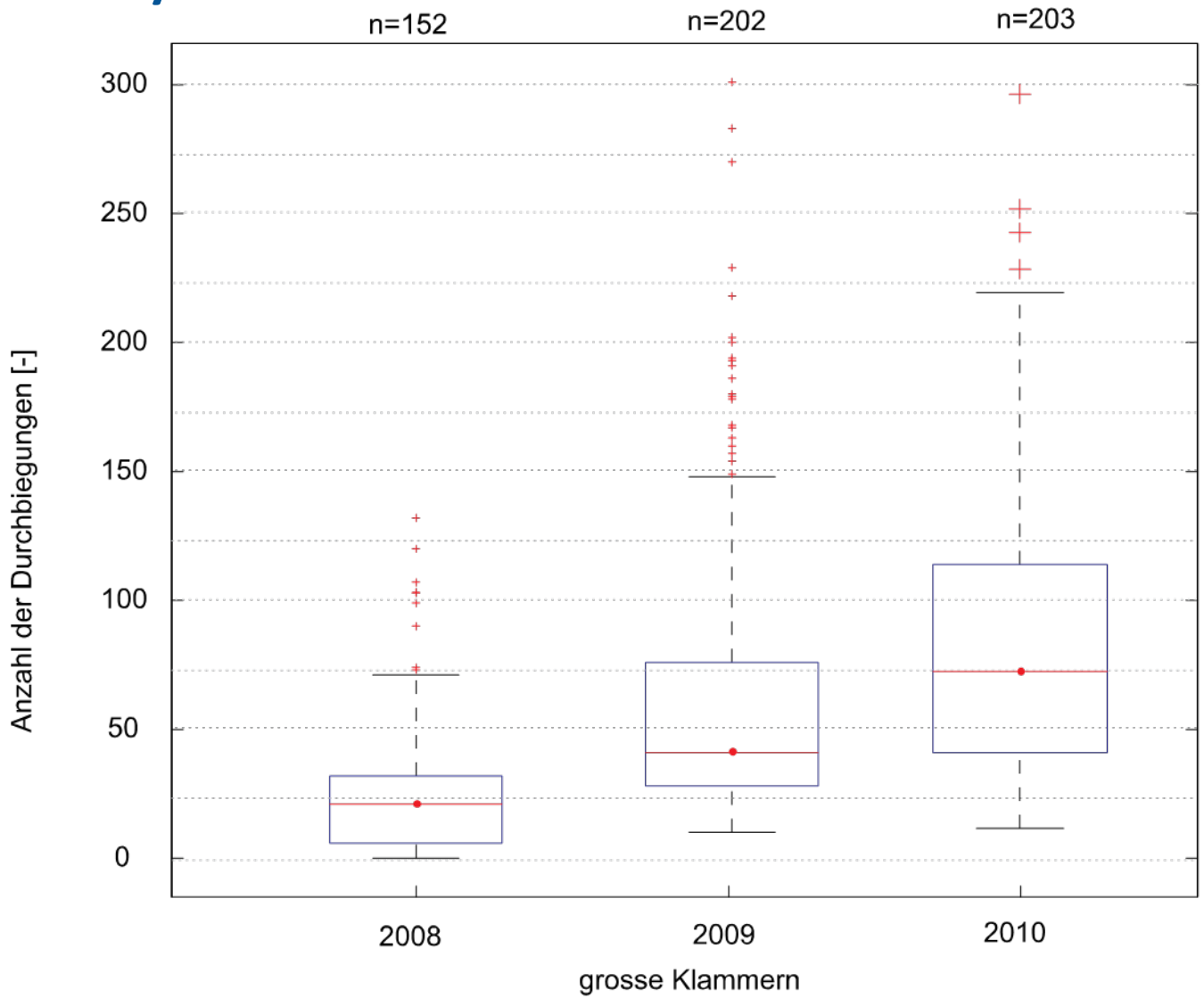
unterer Nachbarschaftswert
kleinste Beobachtung grösser/gleich
unteres Quartil - 1.5 * r

Tukey Box Plot

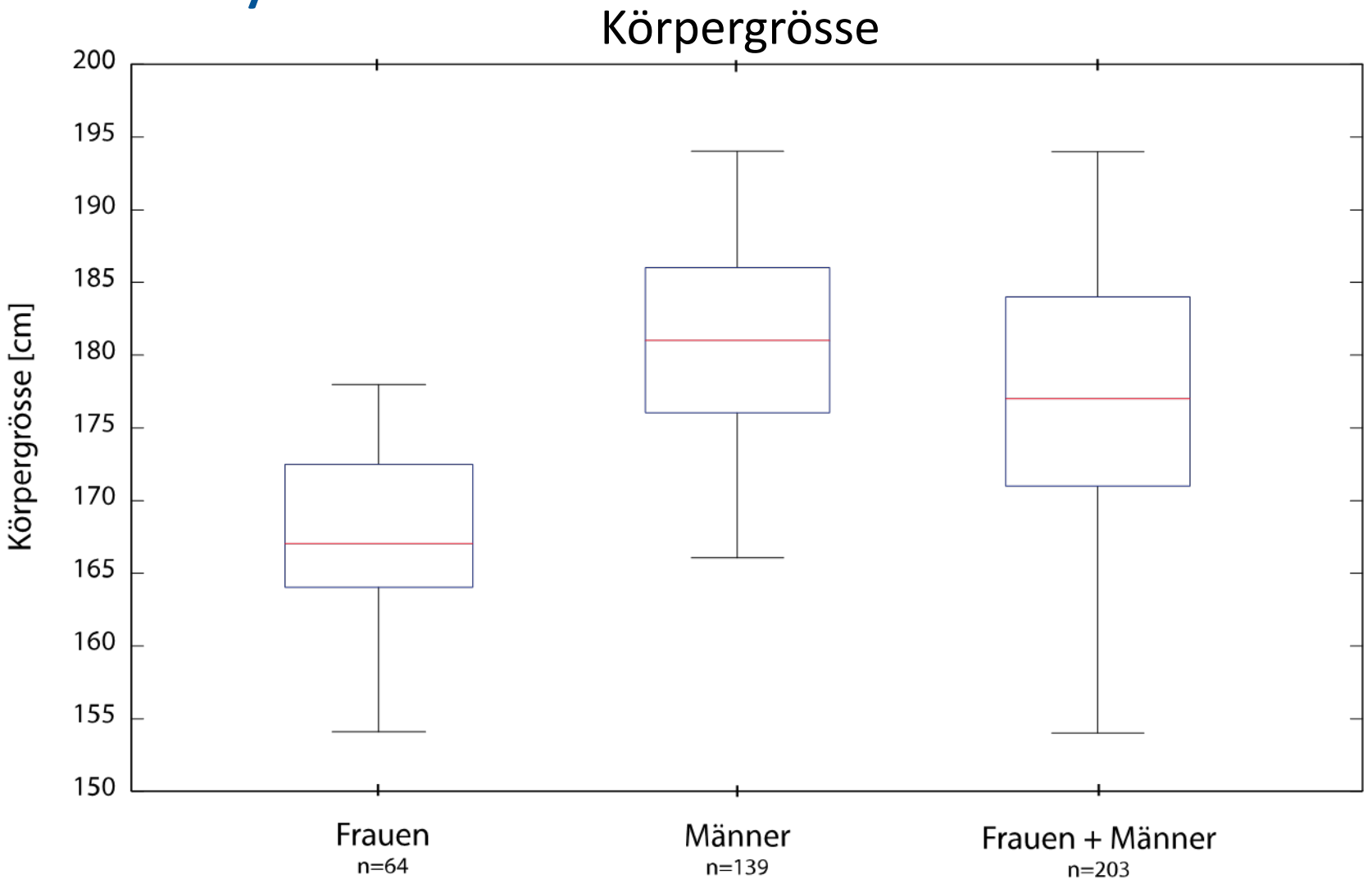


Tukey Box Plot

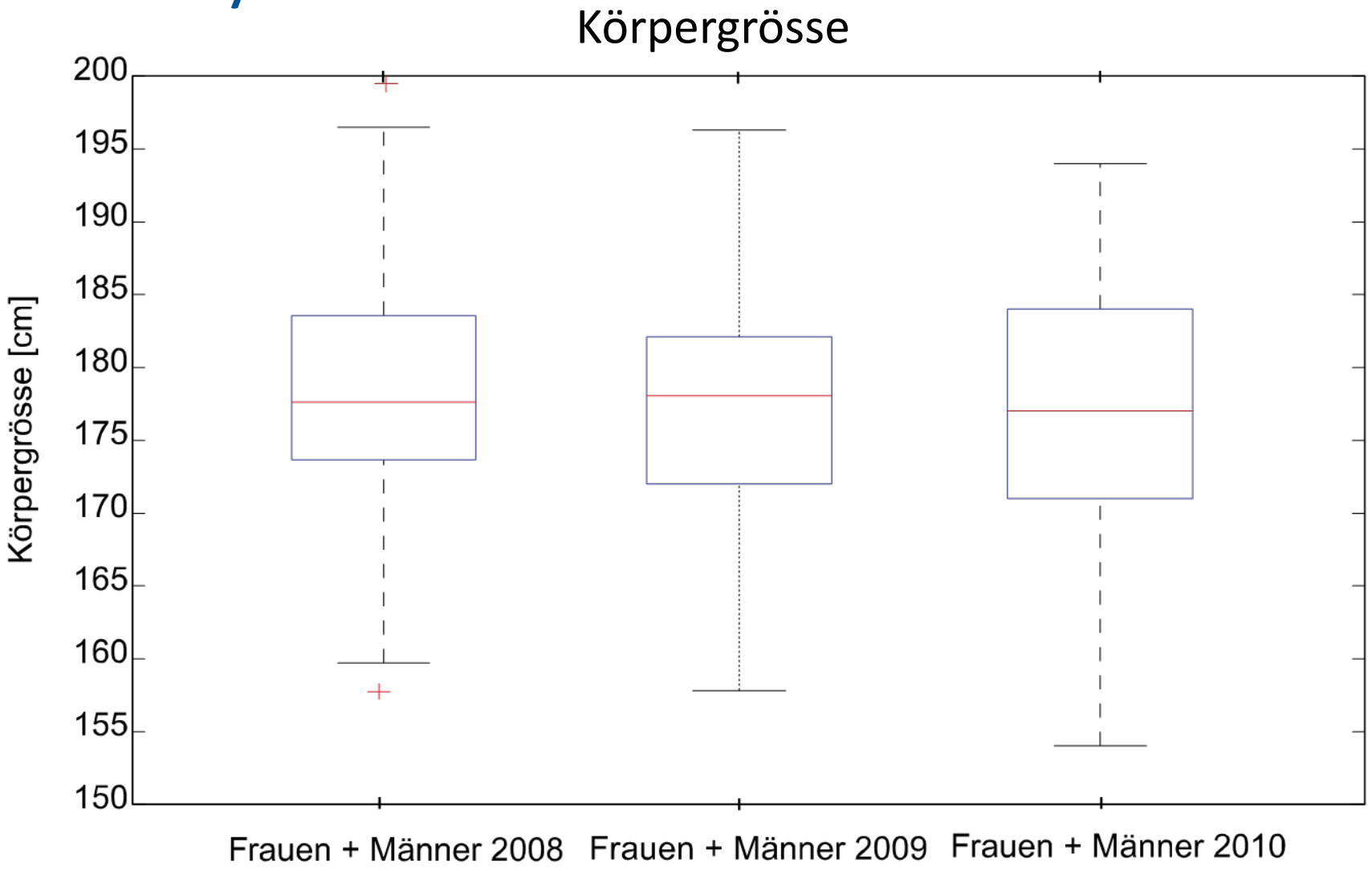
Büroklammern



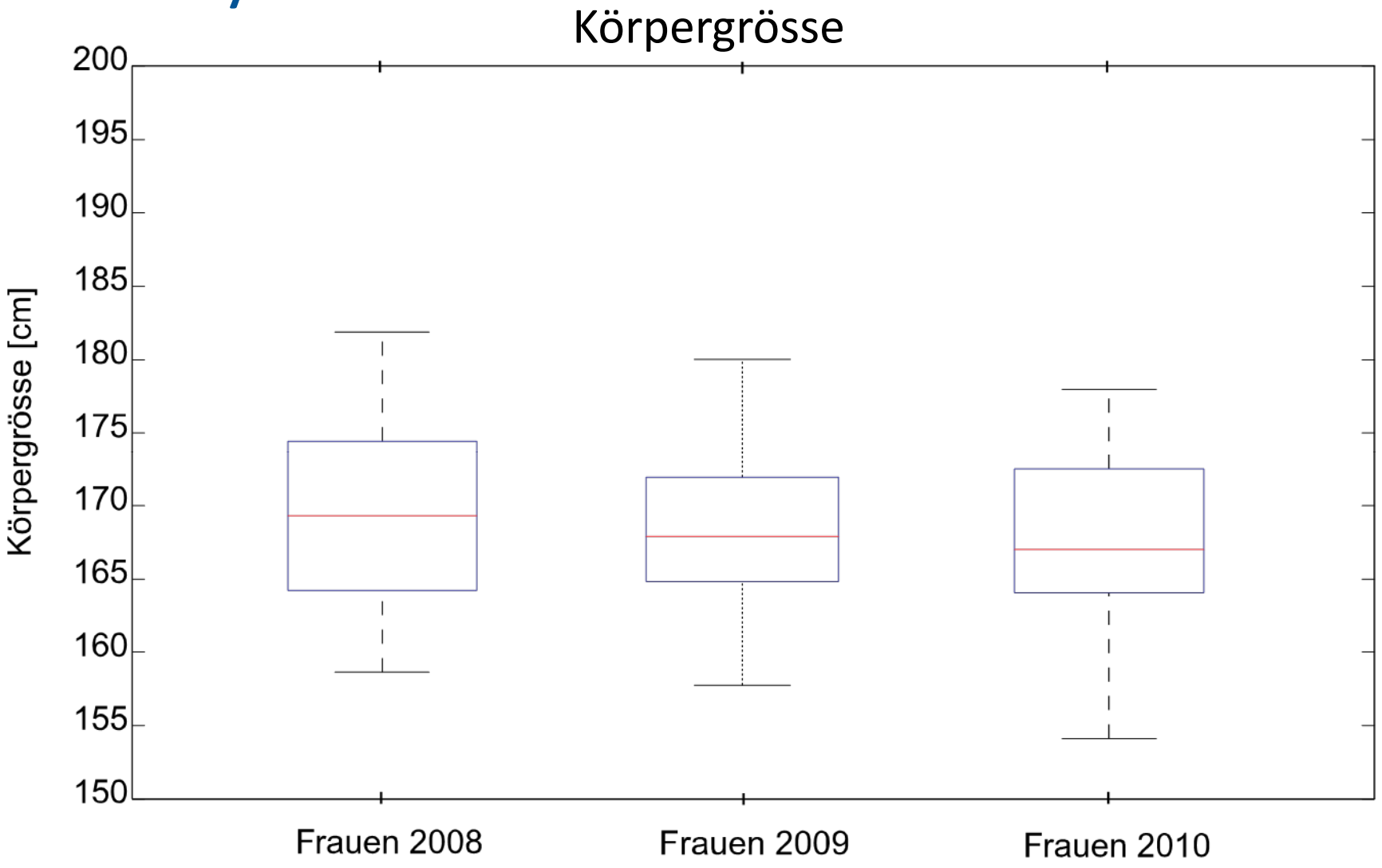
Tukey Box Plot



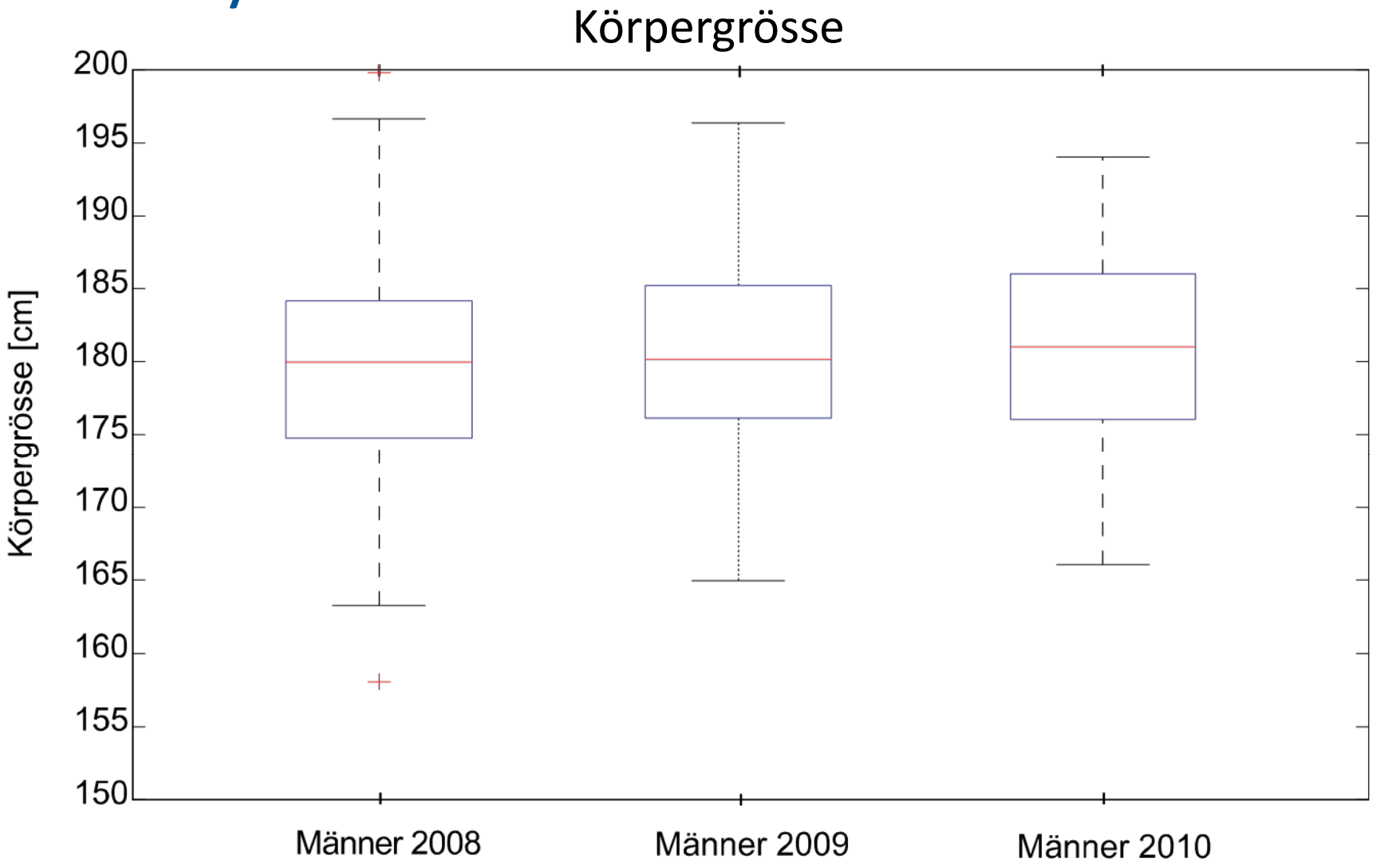
Tukey Box Plot



Tukey Box Plot

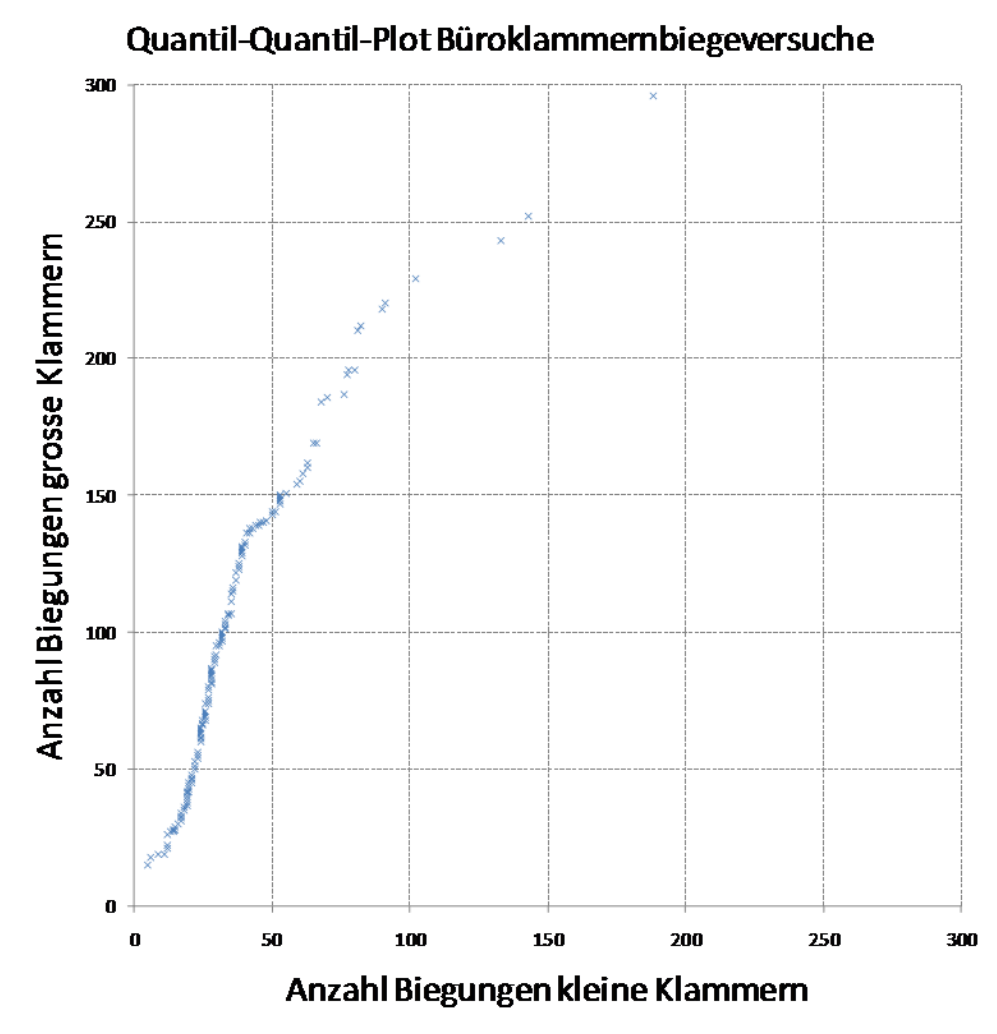


Tukey Box Plot



Q-Q Plots

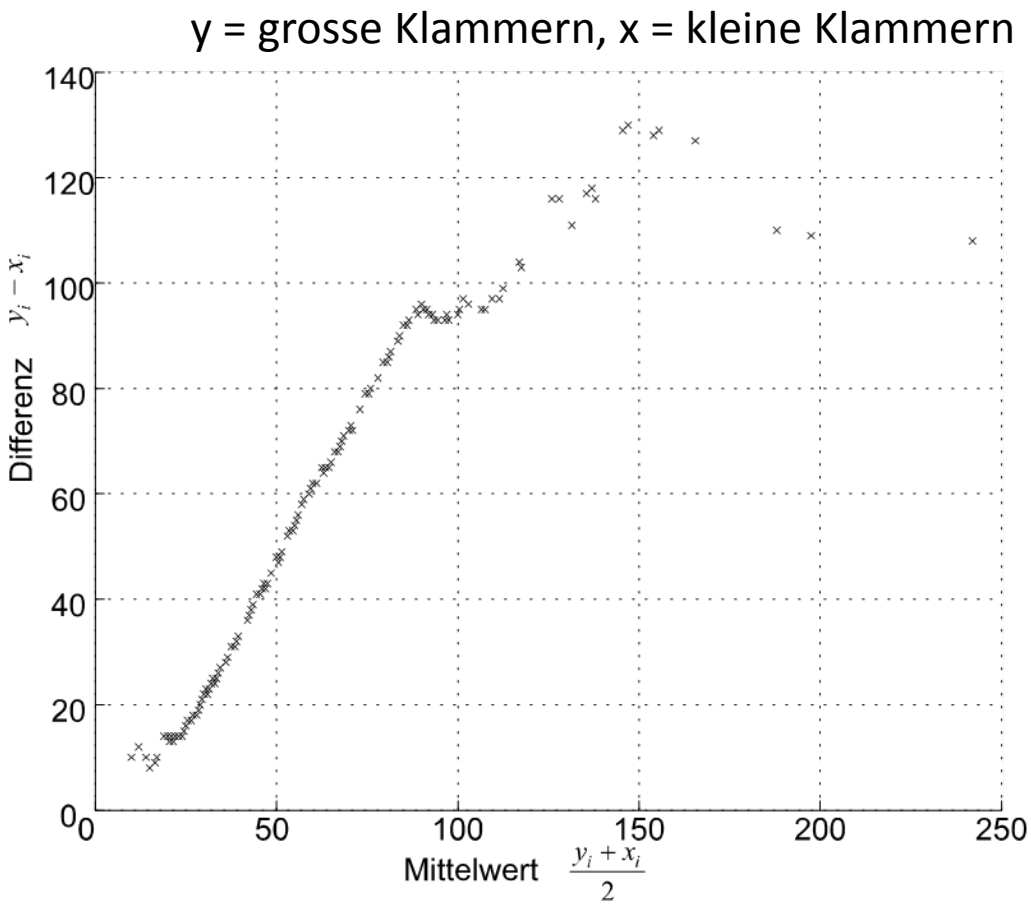
- Q-Q plots dienen zur Darstellung und dem Vergleich von zwei Datenreihen.
- Datenpunkte der beiden Datenreihen mit demselben Quantilwert werden aufgetragen.



Mittelwert-Differenz Plot

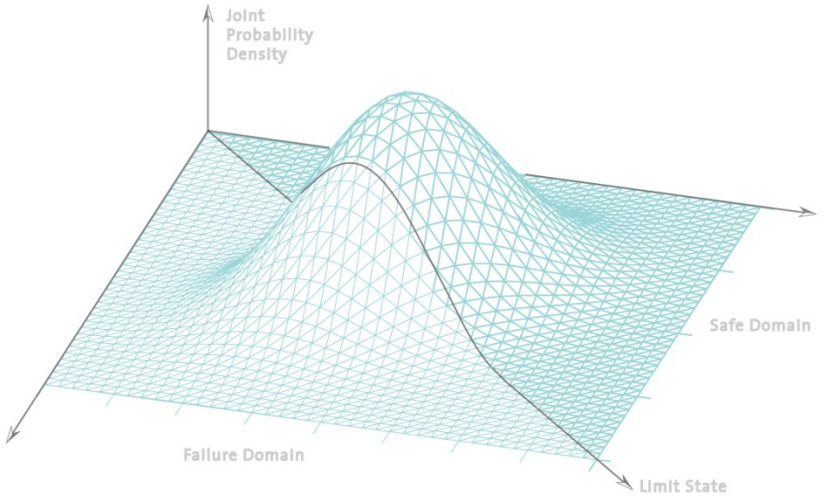
- Mittelwert-Differenz Plots dienen zur Darstellung und dem Vergleich von zwei Datenreihen.

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



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