2. Vorlesung Statistik I
Deskriptive Statistik I

Einführung in die Deskriptive Statistik

Beschreibung von Daten: Deskriptive Statistik

- Liegen Daten vor, ist es häufig zweckmäßig, die Struktur dieser Daten zu beschreiben.

- Beispiele für interessante Fragen: Welcher Zahlenwert kommt am häufigsten vor? Fluktuieren die Zahlenwerte stark?

- Mit dem Begriff „**Deskriptive Statistik**“ fasst man alle Methoden zusammen, die der **Beschreibung von Daten** dienen.

Methoden der Deskriptiven Statistik

Die folgenden Methoden dienen zur Beschreibung von Daten:

- Urlisten und Tabellen
- Häufigkeiten
  - tabellarische Darstellung
  - graphische Darstellung
- Maßzahlen
- Spezielle graphische Darstellungsformen

Datenbeispiel

- In einer bayerischen Mittelschule wurde 2013 die motorische Leistungsfähigkeit von 42 Schüler*innen aus zwei Ganztagsklassen (22 Personen der Klasse 5a und 20 Personen der Klasse 5b) mithilfe des Deutschen Motorik-Tests (DMT) untersucht.

- Die Daten der Schüler*innen in den Untertests „Sprint“ (Schnellkraft), „Anzahl der Liegestütze“ (Kraft) und „Rumpfbeuge“ (Beweglichkeit) werden uns im Folgenden als Beispiel für deskriptivstatistische Berechnungen und graphische Darstellungen dienen.

Urliste

VP	Klasse	Sprint [s]	Anzahl Liegestütz	Rumpfbeuge [cm]
VP1	5a	4,56	12	9
VP2	5a	4,34	20	-1
VP3	5a	4,60	13	-9
VP4	5a	4,54	13	2
VP5	5a	4,22	16	12
VP6	5a	4,29	21	8
VP7	5a	4,18	19	-1
VP8	5a	3,94	15	5
VP9	5a	4,00	17	12
VP10	5a	4,19	20	12
VP11	5a	4,42	16	11
VP12	5a	4,15	15	-4
VP13	5a	4,31	15	-4
VP14	5a	4,53	16	5
VP15	5a	4,50	15	4
⋮		⋮	⋮	⋮
VP36	5b	4,56	19	-9
VP37	5b	4,63	16	4
VP38	5b	4,06	20	-8
VP39	5b	4,90	16	3
VP40	5b	4,98	14	2
VP41	5b	4,43	18	-8
VP42	5b	4,66	16	4

- Die ungeordnete tabellarische Darstellung von Daten nennt man Urliste.
- Meistens sind Urlisten in Tabellenverarbeitungsprogrammen, wie z.B. Excel, gespeichert und werden dann von statistischen Programmen, wie z.B. R, eingelesen.

Zweck von Tabellen

- **Tabellen** dienen dem **Ordnen** und **Zusammenfassen** von Daten.
- Der Wertebereich der **Variablen-** bzw. **Messwerte** x_i mit $i = 1, \dots, n$, wobei *n* die Anzahl der Merkmalsträger bzw. Versuchspersonen ist, ist die Menge aller unterschiedlichen **Messwertausprägungen** x_j, mit $j = 1, \dots, m$, wobei *m* die Anzahl der **aufgetretenen** unterschiedlichen Ausprägungen ist.

- Beispiel:
  - Es wurden die Gewichtsdaten von 5 Personen erhoben. Es liegen daher 5 Messwerte vor. Wie könnte eine Urliste aussehen, die *n* = 5 Messwerte x_i $(x_1, x_2, x_3, x_4, x_5)$ mit *m* = 3 Messwertausprägungen x_j (x_1, x_2, x_3) aufweist?

Beachte: uneindeutige Notation!

- Hinweis: Auch wenn wir hier mit x_j nur die **tatsächlich aufgetretenen** Ausprägungen betrachten, wird es manchmal sinnvoll sein, auch einzelne **nicht aufgetretene, aber mögliche** Ausprägungen zu berücksichtigen. In diesen Fällen ist *m* dann implizit größer als die Anzahl der aufgetretenen Ausprägungen.

Messwerte und Messwertausprägungen

Beispiel für folgende zehn **Messwerte** (Urliste):

Beachte: uneindeutige Notation!

$x_1 = 12$
$x_2 = 20$
$x_3 = 13$
$\vdots$
$x_{10} = 20$

VP	Liegestütz
VP1	12
VP2	20
VP3	13
VP4	13
VP5	16
VP6	21
VP7	19
VP8	15
VP9	17
VP10	20

→

Die Messwerte enthalten acht **Messwertausprägungen**:

$x_1 = 12$
$x_2 = 13$
$x_3 = 15$
$x_4 = 16$
$x_5 = 17$
$x_6 = 19$
$x_7 = 20$
$x_8 = 21$

Liegestütz
12
13
15
16
17
19
20
21

Messwerte x_i beziehen sich auf die Versuchspersonen.

Messwertausprägungen x_j stellen den Wertebereich der aufgetretenen Messwerte x_i dar.

- Als sinnvolle Konvention werden alle aufgetretenen unterschiedlichen Messwertausprägungen **in aufsteigender Reihenfolge** angeordnet.

Anzahl von Messwertausprägungen

- Abhängig von der Anzahl der Messwertausprägungen und der gewünschten Darstellungsform entscheidet man, ob diese in **Kategorien** zusammengefasst oder einzeln aufgeführt werden. Bei der Bildung von **Kategoriengrenzen** ist sinnvollerweise darauf zu achten, dass diese lückenlos den gesamten Messwertebereich abbilden.

VP	Sprint [s]
VP1	4,56
VP2	4,34
VP3	4,60
VP4	4,54
VP5	4,22
VP6	4,29
VP7	4,18
VP8	3,94
VP9	4,00
VP10	4,19
VP11	4,42
VP12	4,15
VP13	4,31
VP14	4,53
VP15	4,50
⋮	
VP38	4,06
VP39	4,90
VP40	4,98
VP41	4,43
VP42	4,66

→

Sprint [s]
3,90 - 4,09
4,10 - 4,29
4,30 - 4,49
4,50 - 4,69
4,70 - 4,89
4,90 - 5,09
5,10 - 5,29
5,30 - 5,49
5,50 - 5,69

VP	Liegestütz
VP1	12
VP2	20
VP3	13
VP4	13
VP5	16
VP6	21
VP7	19
VP8	15
VP9	17
VP10	20
VP11	16
VP12	15
VP13	15
VP14	16
VP15	15
⋮	
VP38	20
VP39	16
VP40	14
VP41	18
VP42	16

→

Liegestütz
9
12
13
14
15
16
17
18
19
20
21

Häufigkeiten

Verschiedene Arten von Häufigkeiten

- Die Zuordnung von Häufigkeiten zu den Messwertausprägungen x_j mit

  $j = 1, \ldots, m$ und stellt eine erste sehr nützliche Form der Datenbeschreibung dar.

- Man unterscheidet **absolute Häufigkeiten**, **relative Häufigkeiten**, **absolute kumulierte Häufigkeiten** und **relative kumulierte Häufigkeiten**.

- Alle Häufigkeitsarten können sowohl in tabellarischer, als auch in graphischer Form dargestellt werden.

absolute Häufigkeit

Anzahl Liegestütz	absolute Häufigkeit *H*
9	1
12	2
13	4
14	3
15	5
16	10
17	2
18	4
19	4
20	5
21	2
Summe	42

Sprint [s]	absolute Häufigkeit *H*
3,90 - 4,09	3
4,10 - 4,29	9
4,30 - 4,49	9
4,50 - 4,69	10
4,70 - 4,89	4
4,90 - 5,09	5
5,10 - 5,29	1
5,30 - 5,49	0
5,50 - 5,69	1
Summe	42

- Die absolute Häufigkeit $H(x_j)$ entspricht für jede Messwertausprägung x_j der **Anzahl** der Merkmalsträger mit dieser Messwertausprägung.

- Abhängig vom inhaltlichen Kontext sowie der Anzahl an möglichen Messwertausprägungen ist es manchmal sinnvoll, auch nicht beobachtete Messwertausprägungen mit $H(x_j) = 0$ zu berücksichtigen.

relative Häufigkeit

Anzahl Liegestütz	absolute Häufigkeit *H*	relative Häufigkeit *h*
9	1	0,024
12	2	0,048
13	4	0,095
14	3	0,071
15	5	0,119
16	10	0,238
17	2	0,048
18	4	0,095
19	4	0,095
20	5	0,119
21	2	0,048
Summe	42	1

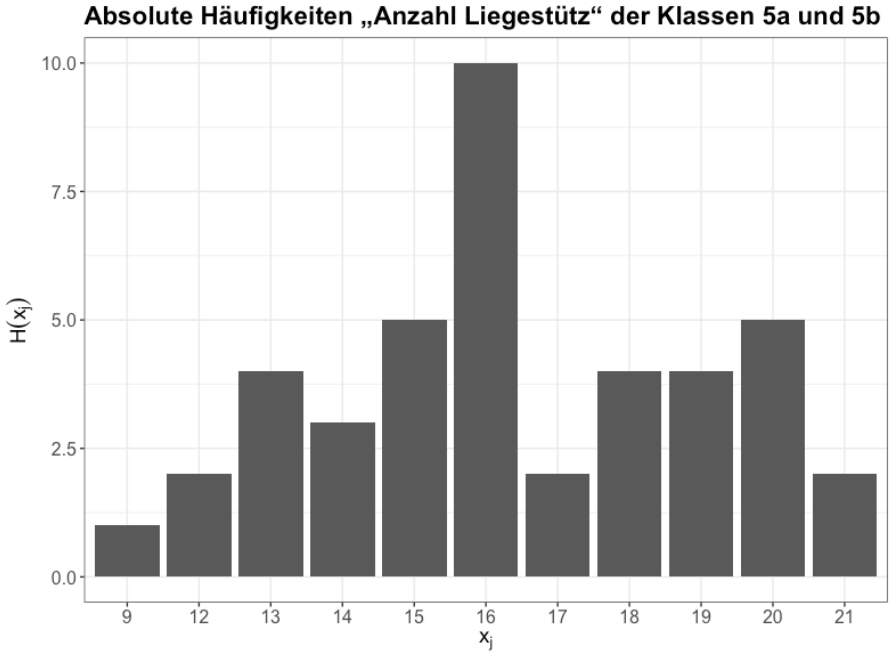
Sprint [s]	absolute Häufigkeit *H*	relative Häufigkeit *h*
3,90 - 4,09	3	0,071
4,10 - 4,29	9	0,214
4,30 - 4,49	9	0,214
4,50 - 4,69	10	0,238
4,70 - 4,89	4	0,095
4,90 - 5,09	5	0,119
5,10 - 5,29	1	0,024
5,30 - 5,49	0	0,000
5,50 - 5,69	1	0,024
Summe	42	1

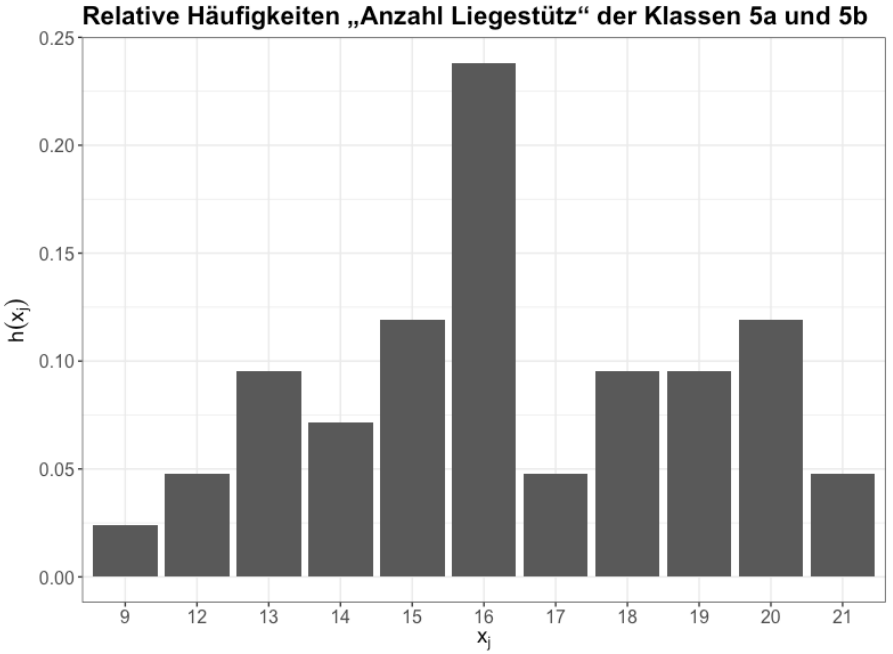
- Die relative Häufigkeit wird gebildet, indem die absolute Häufigkeit durch die Gesamtanzahl *n* der Messwerte bzw. Versuchspersonen geteilt wird:

$$h(x_j) = \frac{H(x_j)}{n}$$

- Relative Häufigkeiten stellen den **Anteil** der Merkmalsträger*innen dar, die die entsprechende Messwertausprägung aufweisen. Sie können auch in Prozent angegeben werden.

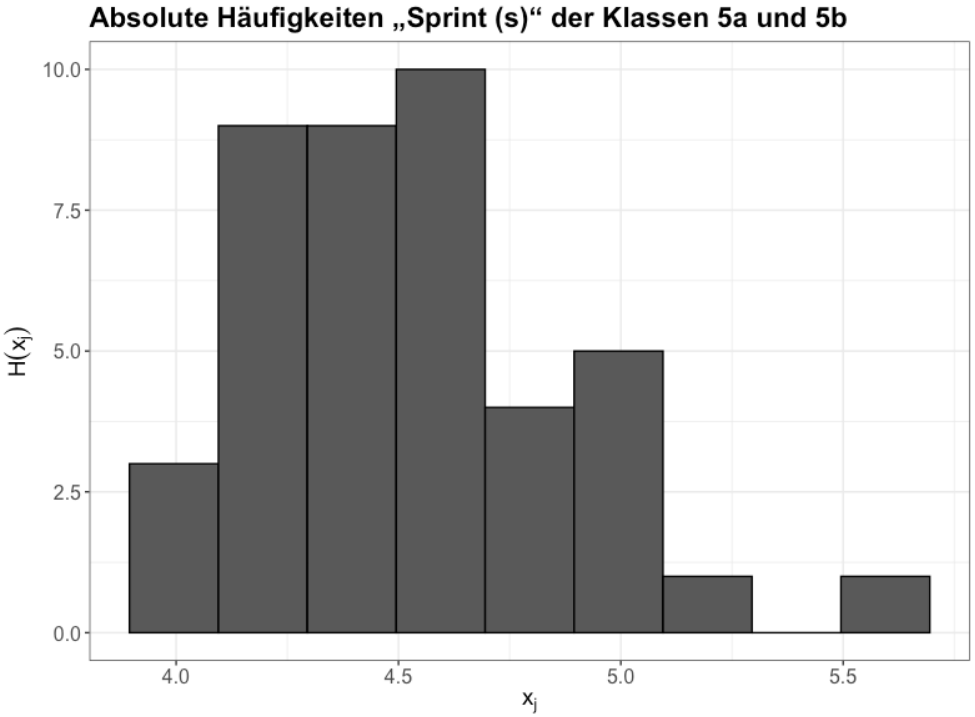
Balkendiagramm (Säulendiagramm)

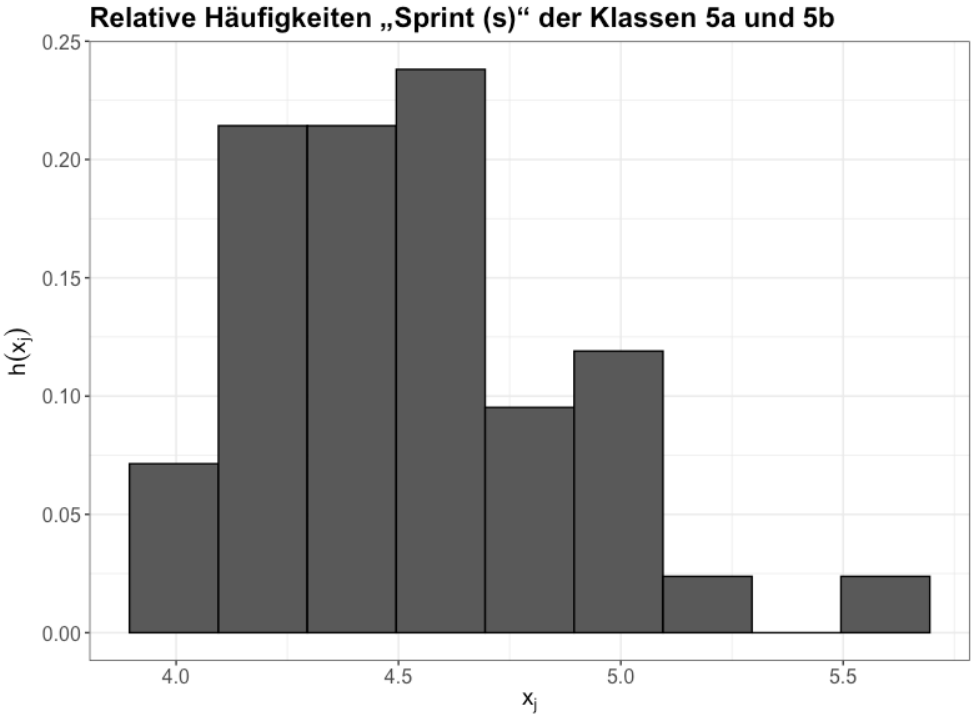




- Balkendiagramme werden zur graphischen Abbildung von Häufigkeiten bei **diskreten Variablen** herangezogen.
- Auf der x-Achse können in Abhängigkeit davon, welche Art der diskreten Variable vorliegt, Zahlen oder Symbole aufgetragen werden.
- Auf der y-Achse sind die jeweiligen Häufigkeiten abgetragen.
- Balkendiagramme sollten auf der y-Achse immer bei null beginnen.

Histogramm





- Histogramme werden zur Abbildung von Häufigkeiten bei **kontinuierlichen Variablen** herangezogen.
- Kontinuität von Variablen wird im Histogramm durch die lückenlose Aneinanderreihung der Balken veranschaulicht. Die Balkenbreite ergibt sich aus der Breite der gebildeten Kategorien.

Exkurs: Summen

- Betrachtet werden *n* reelle Zahlen $a_1, a_2, \ldots, a_n$. Die Summe der Zahlen notiert man folgendermaßen:

  $a_1 + a_2 + \ldots + a_n = \sum_{i=1}^{n} a_i$, mit $n \in \mathbb{N}$,

  i heißt hierbei **Laufindex**.

- Das **Summenzeichen** $\sum$ vereinfacht die Schreibweise von Summen und wird in der Statistik häufig verwendet.

- Es gibt Rechenregeln für Summen. Drei häufig verwendete Rechenregeln:

$$\sum_{i=1}^{n} a = n \cdot a$$

$$\sum_{i=1}^{n} c \cdot a_i = c \cdot \sum_{i=1}^{n} a_i$$

$$\sum_{i=1}^{n} (a_i \pm b_i) = \sum_{i=1}^{n} a_i \pm \sum_{i=1}^{n} b_i$$

Absolute kumulierte Häufigkeit I

Anzahl Liegestütz	absolute Häufigkeit H	kumulierte absolute Häufigkeit H_{kum}
9	1	1
12	2	3
13	4	7
14	3	10
15	5	15
16	10	25
17	2	27
18	4	31
19	4	35
20	5	40
21	2	42

Für die absolute kumulierte Häufigkeit einer Messwertausprägung (bei einer mindestens ordinal skalierten Variable) wird die absolute Häufigkeiten dieser Messwertausprägung und aller kleineren Messwertausprägungen aufsummiert:

$$H_{kum}(x_k) = \sum_{j=1}^{k} H(x_j)$$

Absolute kumulierte Häufigkeit II

Anzahl Liegestütz	absolute Häufigkeit *H*	kumulierte absolute Häufigkeit H_{kum}
9	1	1
12	2	3
13	4	7
14	3	10
15	5	15
16	10	25
17	2	27
18	4	31
19	4	35
20	5	40
21	2	42

Beispiel für $k = 4$:

$$H_{kum}(x_4) = \sum_{j=1}^{4} H(x_j) = H(x_1) + H(x_2) + H(x_3) + H(x_4) = 1 + 2 + 4 + 3 = 10$$

Absolute kumulierte Häufigkeit III

Anzahl Liegestütz	absolute Häufigkeit *H*	kumulierte absolute Häufigkeit H_{kum}
9	1	1
12	2	3
13	4	7
14	3	10
15	5	15
16	10	25
17	2	27
18	4	31
19	4	35
20	5	40
21	2	42

- Für die absolute kumulierte Häufigkeit einer Messwertausprägung (bei einer mindestens ordinal skalierten Variable) wird die absolute Häufigkeiten dieser Messwertausprägung und aller kleineren Messwertausprägungen aufsummiert:

$$H_{kum}(x_k) = \sum_{j=1}^{k} H(x_j)$$

- Die absolute kumulierte Häufigkeit einer Messwertausprägung x_k entspricht der **Anzahl** der Merkmalsträger*innen, die die Messwertausprägung x_k oder eine kleinere Messwertausprägung aufweisen.

Relative kumulierte Häufigkeit

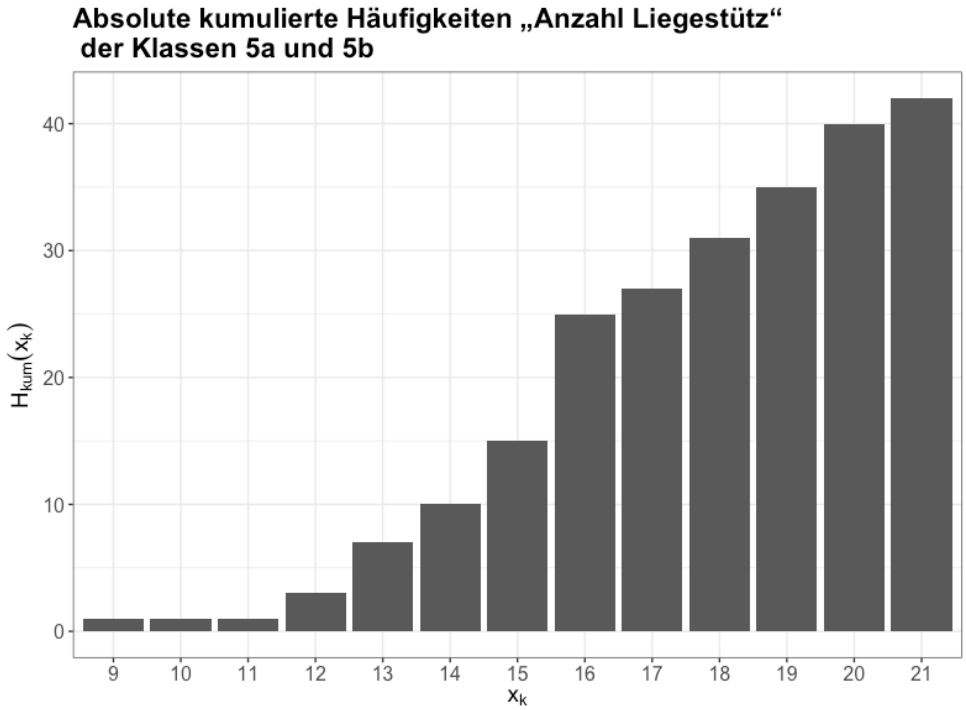
Anzahl Liegestütz	relative Häufigkeit *h*	kumulierte relative Häufigkeit h_{kum}	kumulierte relative Häufigkeit h_{kum} in %
9	0,024	0,024	2,38
12	0,048	0,071	7,14
13	0,095	0,167	16,67
14	0,071	0,238	23,81
15	0,119	0,357	35,71
16	0,238	0,595	59,52
17	0,048	0,643	64,29
18	0,095	0,738	73,81
19	0,095	0,833	83,33
20	0,119	0,952	95,24
21	0,048	1,000	100,00

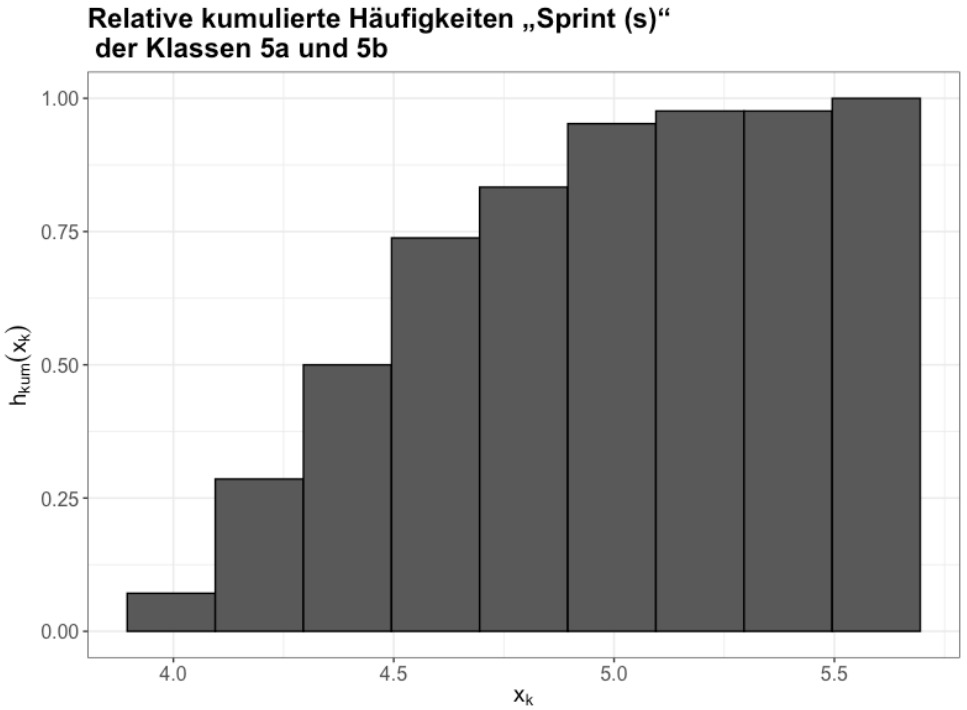
- Die relative kumulierte Häufigkeit wird gebildet, indem die absolute kumulierte Häufigkeit durch die Gesamtanzahl *n* der Messwerte bzw. Versuchspersonen geteilt wird:

$$h_{kum}(x_k) = \frac{\sum_{j=1}^{k} H(x_j)}{n} = \frac{H_{kum}(x_k)}{n}$$

- Die relative kumulierte Häufigkeit einer Messwertausprägung x_k entspricht dem **Anteil** der Merkmalsträger*innen, die die Messwertausprägung x_k oder eine kleinere Messwertausprägung aufweisen.

Graphische Darstellung von kumulierten Häufigkeiten





- Auch absolute kumulierte und relative kumulierte Häufigkeiten können mit einem entsprechenden Balkendiagramm (bei diskreten Variablen) oder Histogramm (bei kontinuierlichen Variablen) dargestellt werden.
- Dabei ist es fast immer sinnvoll, auch nicht aufgetretene Messwertausprägungen zu berücksichtigen (dies können wir erst später vollständig wertschätzen).

Das „Lesen“ von Häufigkeitstabellen

Klassen 5a und 5b			
Rumpfbeuge [cm]	*H*	*h* (%)	h_{kum} (%)
-15	1	2,4	2,4
-10	1	2,4	4,8
- 9	2	4,8	9,5
- 8	2	4,8	14,3
-7	1	2,4	16,7
-5	1	2,4	19,0
-4	3	7,1	26,2
-2	1	2,4	28,6
-1	5	11,9	40,5
0	1	2,4	42,9
1	1	2,4	45,2
2	4	9,5	54,8
3	3	7,1	61,9
4	5	11,9	73,8
5	2	4,8	78,6
6	1	2,4	81,0
8	2	4,8	85,7
9	1	2,4	88,1
11	1	2,4	90,5
12	3	7,1	97,6
13	1	2,4	100,0

Es lassen sich relevante Informationen über die Verteilung der Messwerte aus einer Häufigkeitstabelle „herauslesen“, z.B.:

- Welcher Wert kommt am häufigsten vor?
- Sind außergewöhnliche Werte gemessen worden?
- Gibt es nur einen oder gibt es mehrere Werte, die sehr häufig vorkommen?

Deskriptive Statistik

- Bislang:
  - Häufigkeiten

- Jetzt:
  - Maßzahlen

Maßzahlen

Maßzahlen

- Maßzahlen drücken die Eigenschaften von mehreren Messwerten in komprimierter Form durch einen numerischen Wert aus.

- Maßzahlen lassen sich aufgliedern in:
  - Maße der zentralen Tendenz (Synonym: Lagemaße)
  - Streuungsmaße (Synonym: Dispersionsmaße)

Maße der zentralen Tendenz

- Mithilfe der **Maße der zentralen Tendenz** wird ein Wert ermittelt, **der besonders „typisch“ oder „kennzeichnend“** für die erhobenen Messwerte ist.

- Unter dem Begriff Maße der zentralen Tendenz wird eine Reihe von Maßen zusammengefasst. Die in den Sozialwissenschaften am häufigsten verwendeten Lagemaße sind:
  - Modalwert (Synonym: Modus)
  - arithmetisches Mittel (Synonym: Mittelwert)
  - Median
  - Quantile

Maße der zentralen Tendenz: Modalwert (Synonym: Modus)

	H	h	h_{kum} (%)
trifft gar nicht zu	15	0,15	15
trifft weniger zu	22	0,22	37
trifft zu	35	0,35	72
trifft mehr zu	12	0,12	84
trifft sehr zu	16	0,16	100

Sprint [s]	absolute Häufigkeit H
3,90 - 4,09	3
4,10 - 4,29	9
4,30 - 4,49	9
4,50 - 4,69	10
4,70 - 4,89	4
4,90 - 5,09	5
5,10 - 5,29	1
5,30 - 5,49	0
5,50 - 5,69	1

- Unter dem **Modalwert** versteht man **die Messwertausprägung**, die am häufigsten beobachtet wurde.
- Wo liegt der Modalwert der nebenstehenden Messwerte?
- Hinweis: Es kann mehr als einen Modalwert geben, falls mehrere Messwertausprägungen die gleiche maximale Häufigkeit aufweisen.

Maße der zentralen Tendenz: arithmetisches Mittel

- Das **arithmetische Mittel** ist die Summe aller aufgetretenen Messwerte geteilt durch die Anzahl der Versuchspersonen:

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

- Interpretation: Durchschnittlicher Messwert.
- Beispiel:

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}
2	1	1	1	4	6	3	3	5	5	3	2	2	3	3

$$\bar{x} = \frac{1}{15}\sum_{i=1}^{15} x_i = \frac{2+1+1+1+4+6+3+3+5+5+3+2+2+3+3}{15} = 2{,}93$$

Maße der zentralen Tendenz: Median

- Der **Median** ist der Wert, der eine geordnete Messwertreihe in zwei gleichgroße Hälften aufteilt: Er wird bestimmt indem die mittlere Position der geordneten Urliste ermittelt wird.

- Interpretation: (Mindestens) 50% der Merkmalsträger*innen haben einen Messwert, der kleiner oder gleich dem Median ist.

Berechnung des Medians

- Zur Berechnung des Medians ist es notwendig, dass **alle Messwerte** x_i mit $i = 1, \dots, n$ einer Urliste in **aufsteigender Reihenfolge** geordnet werden:

  $$x_{(1)} \leq x_{(2)} \leq \dots \; x_{(n)}$$

- Hinweis: Das ist natürlich erst dann sinnvoll, wenn mindestens Ordinalskalenniveau der Variable vorliegt und eine Reihenfolge der Werte überhaupt möglich ist.

- Bemerkung: x_1 bezeichnet den Messwert der VP 1 einer ungeordneten Urliste, während $x_{(1)}$ den kleinsten Messwert einer geordneten Urliste bezeichnet.

- Rechenvorschrift für die Berechnung des Medians:

$$Md = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & falls\ n\ ungerade \\ \dfrac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2} & falls\ n\ gerade \end{cases}$$

- Hinweis: Es gibt noch eine Reihe von alternativen Methoden, um den Median zu berechnen. Diese werden wir hier nicht besprechen.

Berechnung des Medians: Beispiel 1 (für *n* gerade)

- Urliste von Noten:

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}
2	1	1	1	4	6	3	3	5	5	3	4	2	4	4	5

- geordnete Urliste von Noten:

$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$	$x_{(6)}$	$x_{(7)}$	$x_{(8)}$	$x_{(9)}$	$x_{(10)}$	$x_{(11)}$	$x_{(12)}$	$x_{(13)}$	$x_{(14)}$	$x_{(15)}$	$x_{(16)}$
1	1	1	2	2	3	3	3	4	4	4	4	5	5	5	6

- Berechnung des Medians (für $n = 16$):

$$Md = \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2} = \frac{x_{\left(\frac{16}{2}\right)} + x_{\left(\frac{16}{2}+1\right)}}{2} = \frac{x_{(8)} + x_{(9)}}{2} = \frac{3+4}{2} = 3{,}5$$

Berechnung des Medians: Beispiel 2 (für *n* ungerade)

- Urliste von Noten:

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}
2	1	1	1	4	6	3	3	5	5	3	2	2	3	3

- geordnete Urliste von Noten:

$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$	$x_{(6)}$	$x_{(7)}$	$x_{(8)}$	$x_{(9)}$	$x_{(10)}$	$x_{(11)}$	$x_{(12)}$	$x_{(13)}$	$x_{(14)}$	$x_{(15)}$
1	1	1	2	2	2	3	3	3	3	3	4	5	5	6

- Berechnung des Medians (für *n* = 15):

$$Md = x_{\left(\frac{n+1}{2}\right)} = x_{\left(\frac{15+1}{2}\right)} = x_{(8)} = 3$$

Gegenüberstellung von Mittelwert und Median

- Mittelwert:
  - Vorteil: In die Berechnung fließt die Information aller Messwerte ein.
  - Nachteil: Der Mittelwert reagiert empfindlich auf Ausreißerwerte.

- Median:
  - Vorteil: Der Median ist nicht ausreißersensitiv.
  - Nachteil: Da nur die Messwerte in der Mitte der geordneten Urliste in die Berechnung des Medians eingehen, wird Information „verschenkt“. Der Median ist informationsärmer als der Mittelwert.

p-%-Quantile

- Wir erinnern uns an die Interpretation des Medians: (Mindestens) **50%** der Merkmalsträger*innen haben einen Messwert, der kleiner oder gleich dem **Median** ist.
- Zur Definition des *p*-%-Quantils verallgemeinern wir diese Eigenschaft nun auf beliebige Prozentanteile *p.*
- (Mindestens) ***p* %** der Merkmalsträger*innen haben einen Messwert, der kleiner oder gleich dem ***p*-%-Quantil** ist.
- Der Median entspricht somit dem 50%-Quantil.
- Für die konkrete Berechnung der Quantile gibt es ebenfalls verschiedene Methoden, deren Ergebnisse sich manchmal leicht unterscheiden. Mit diesen Feinheiten werden wir uns hier nicht genauer beschäftigen. Prinzipiell funktioniert die Berechnung ähnlich wie beim Median.

Perzentile, Dezile und Quartile

- Häufig betrachtet man bestimmte Quantile: **Perzentile, Dezile** und **Quartile**.
  - Perzentile : $p = 1,\ 2,\ \ldots,\ 99$
  - Dezile : $p = 10,\ 20, \ldots, 90$
  - Quartile : $p = 25,\ 50,\ 75$

Quartile am Beispiel

Sprint (s):

p-%	0	25	50	75	100
Quantil	3.94	4.28	4.47	4.79	5.59

- Das erste Quartil (p = 25) ist $Q_1 = 4.28$. Also haben 25% der Schüler*innen beim Sprint einen Wert kleiner oder gleich 4.28 s erreicht.

- Das zweite Quartil (p = 50) ist $Q_2 = 4.47$. Also haben 50% der Schüler*innen beim Sprint einen Wert kleiner oder gleich 4.47 s erreicht.

- Das dritte Quartil (p = 75) ist $Q_3 = 4.79$. Also haben 75% der Schüler*innen beim Sprint einen Wert kleiner oder gleich 4.79 s erreicht.

Streuungsmaße (Synonym: Dispersionsmaße)

- Unter dem Begriff **Streuungsmaße** werden zahlreiche Maße zusammengefasst, die das Ausmaß der Unterschiedlichkeit der Messwerte quantifizieren.
- Die bekanntesten und – im statistischen Sinne – wichtigsten Streuungsmaße sind die **Varianz** und die **Standardabweichung**.
- Darüber hinaus werden wir noch den **Interquartilsabstand** besprechen.

(empirische) Varianz

- Die (empirische) **Varianz s^2_{emp}** ist definiert als:

$$s^2_{emp} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- Die Streuung der Messwerte wird quantifiziert, indem die Abweichungen der Messwerte vom Mittelwert quadriert, aufsummiert und anschließend gemittelt werden.
- Je größer die Varianz, desto mehr streuen die Messwerte um das arithmetische Mittel.

(empirische) Standardabweichung

- Die **Standardabweichung s_{emp}** ist die (positive) Quadratwurzel aus der Varianz:

$$s_{emp} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- Varianz und Standardabweichung sind statistisch gesehen die wichtigsten Streuungsmaße. Sie haben jedoch keine besonders intuitive Interpretation.

Interquartilsabstand

- Der **Interquartilbereich** ist definiert als der Bereich der Werte zwischen erstem und drittem Quartil:

$$IQB = [Q_1, Q_3]$$

- In diesem Bereich liegen die mittleren 50 Prozent der Werte.

Quartile „Sprint (s)“:

p-%	0	25	50	75	100
Quantil	3.94	4.28	4.47	4.79	5.59

$$IQB = [4{,}28;\ 4{,}79]$$

- Der **Interquartilsabstand** (IQA) ist die Differenz von drittem und ersten Quartil:

$$IQA = Q_3 - Q_1$$

- Der Interquartilsabstand für „Sprint (s)“ ist folglich:

$$IQA = 4{,}79 - 4{,}28 = 0{,}51$$

Deskriptive Statistik

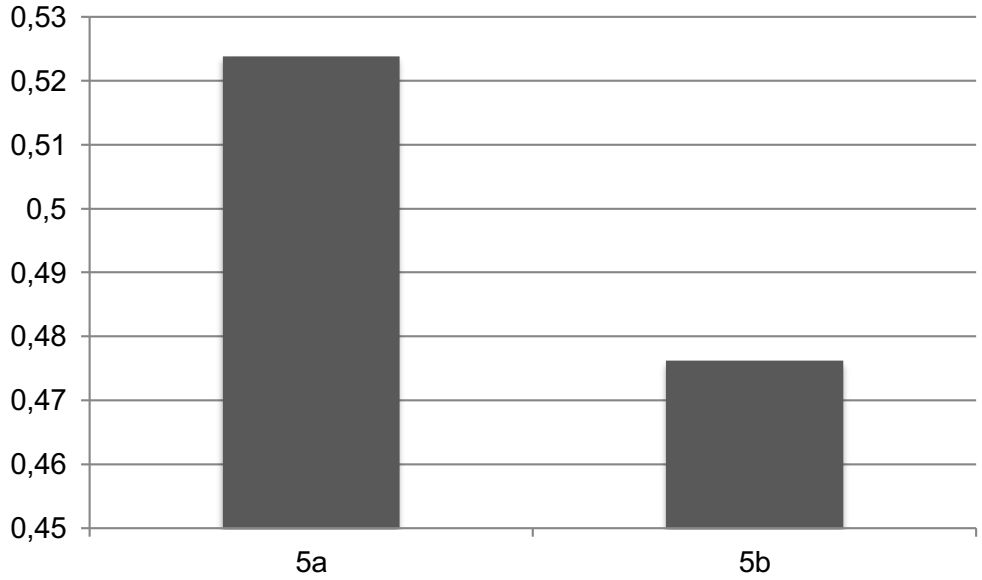
- Bislang:
  - Häufigkeiten
  - Maßzahlen

- Jetzt:
  - Spezielle graphische Darstellungsformen

Spezielle graphische Darstellungsformen

allgemeine Hinweise zu graphischen Darstellungen

- Bsp. für eine nicht optimale graphische Darstellung von relativen Häufigkeiten:



- Bei graphischen Darstellungen ist darauf zu achten, dass die gewählte **Skalierung der y - Achse** das Größenverhältnis der Balken so abbildet, dass kein falscher Eindruck entsteht.
- **Achsenbeschriftungen** sind unerlässlich.
- **Diagrammtitel** sind hilfreich.

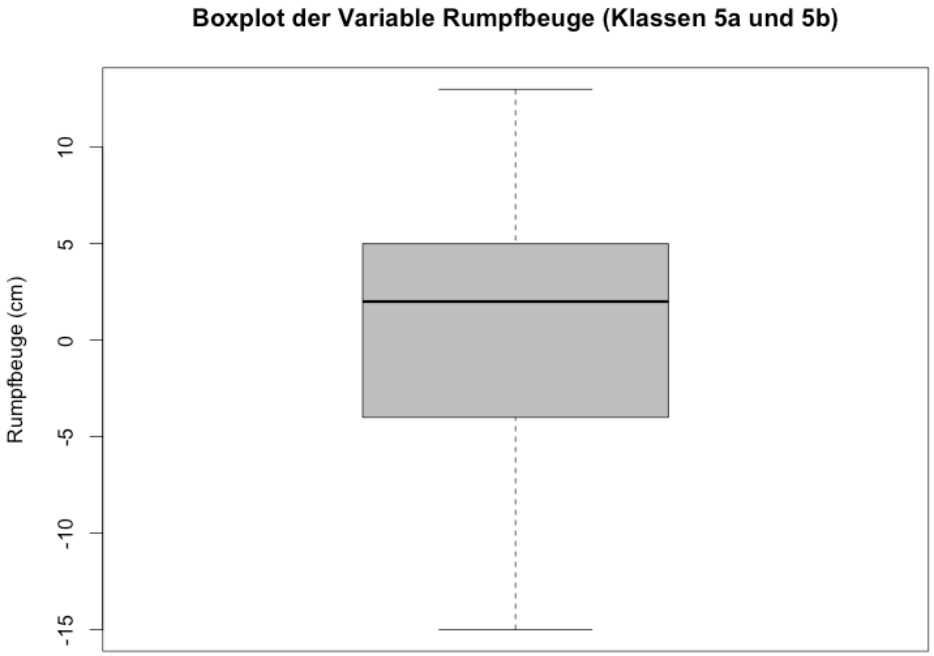
graphischen Darstellungen in der deskriptiven Statistik

- Bei der graphischen Darstellung von Daten im Rahmen der Deskriptivstatistik gibt es außer einigen sinnvollen Konventionen keine allgemeingültigen Regeln, was eine „gute“ graphische Darstellung ausmacht.

- In jedem konkretem Anwendungsfall gibt es sehr viele Möglichkeiten, nützliche Informationen in den Daten so zu visualisieren, dass die Graphik leicht verständlich ist, ohne gleichzeitig wichtige Details zu unterschlagen oder den Betrachter auf andere Weise in die Irre zu führen.
  - Praktische Übung in den angewandten Lehrveranstaltungen (z.B. Empra)

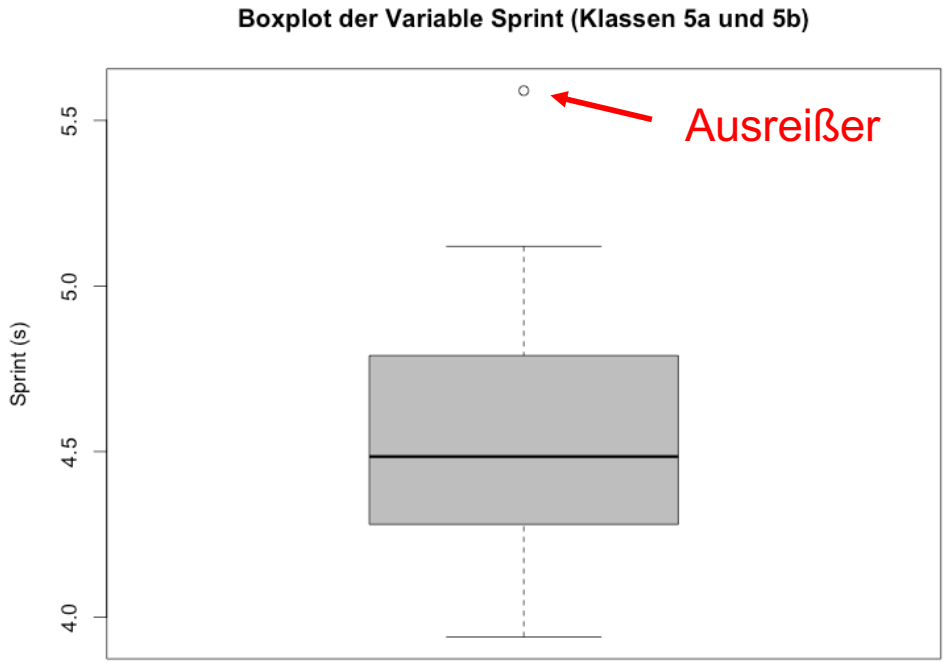
- Trotzdem gibt es einige typische Darstellungsformen, für die sich in der Deskriptivstatistik Standards etabliert haben und die deshalb eigene Namen besitzen:
  - Wichtiges Beispiel: der **Box-Plot**

Box-Plot



- Box-Plots werden vor allem zur Abbildung von **metrischen** Variablen herangezogen. Aber auch kategoriale Variablen, die Ranginformation enthalten, können mithilfe eines Box-Plots graphisch dargestellt werden.
- Bei Box-Plots wird die Struktur der Daten im Wesentlichen durch **Hervorhebung von fünf Werten** dargestellt: Minimaler und maximaler Messwert, sowie die Quartile Q_1, Q_2 und Q_3. Die Länge der Box entspricht somit dem Interquartilsabstand.
- Ein Box-Plot komprimiert also die in einem Histogramm enthaltene Information.

Box-Plot



- Darüber hinaus werden in Box-Plots außergewöhnlich kleine bzw. große Messwerte, so genannte **Ausreißerwerte** gekennzeichnet.
- Es gibt – je nach statistischer Software – leicht abweichende Konventionen, ab wann ein Messwert als Ausreißer qualifiziert wird.
- Falls Ausreißer vorliegen, stellen die Enden der gestrichelten Linien nicht mehr den minimalen bzw. maximalen Messwert dar, sondern nur noch den minimalen bzw. maximalen Messwert nach Entfernung der Ausreißer.

Zusammenfassung

Was Sie aus dieser VL mitnehmen

- Deskriptive Statistik umfasst Methoden, die der **Beschreibung** von Daten dienen.
- Wir unterscheiden die **Messwerte** (x_i), der einzelnen untersuchten Einheiten (z.B. Personen) von den **Messwertausprägungen** (x_j), die diese Messwerte annehmen können.
- Die **Häufigkeiten** von Messwertausprägungen können wir als absolute, relative, absolute kumulierte und relative kumulierte Häufigkeiten zusammenfassen.
- Typische Werte in Daten werden mit **Lagemaßen** (Mittelwert, Median, Modus, Quantile) beschrieben, die Unterschiedlichkeit von Daten mit **Streuungsmaßen** (Varianz, Standardabweichung, Interquartilsabstand).
- Mit **Balkendiagrammen** oder **Histogrammen** lassen sich Häufigkeiten übersichtlich darstellen. Ein **Boxplot** veranschaulicht sowohl Lagemaße wie die drei Quartile Median, Q_1 und Q_3, als auch den Interquartilsabstand und den gesamten Bereich beobachteter Werte (Range).