Statistik

Corona - Notversion

M. Oettinger

20.04.2020

Stichprobe

Nachschlag: die Stichprobe zum Kurs: Alter, Größe und Haarfarbe
Die Urliste für das Merkmal Alter ist

$$\{x_i\} = (19, 19, 29, 20, 20, 20, 20, 33, 21, 23, 22, 22, 19, 23, 24, 19, 20, 21, 21, 21, 19, 22, 19, 48)$$

Der Umfang der Stichprobe ist $n = 24$.

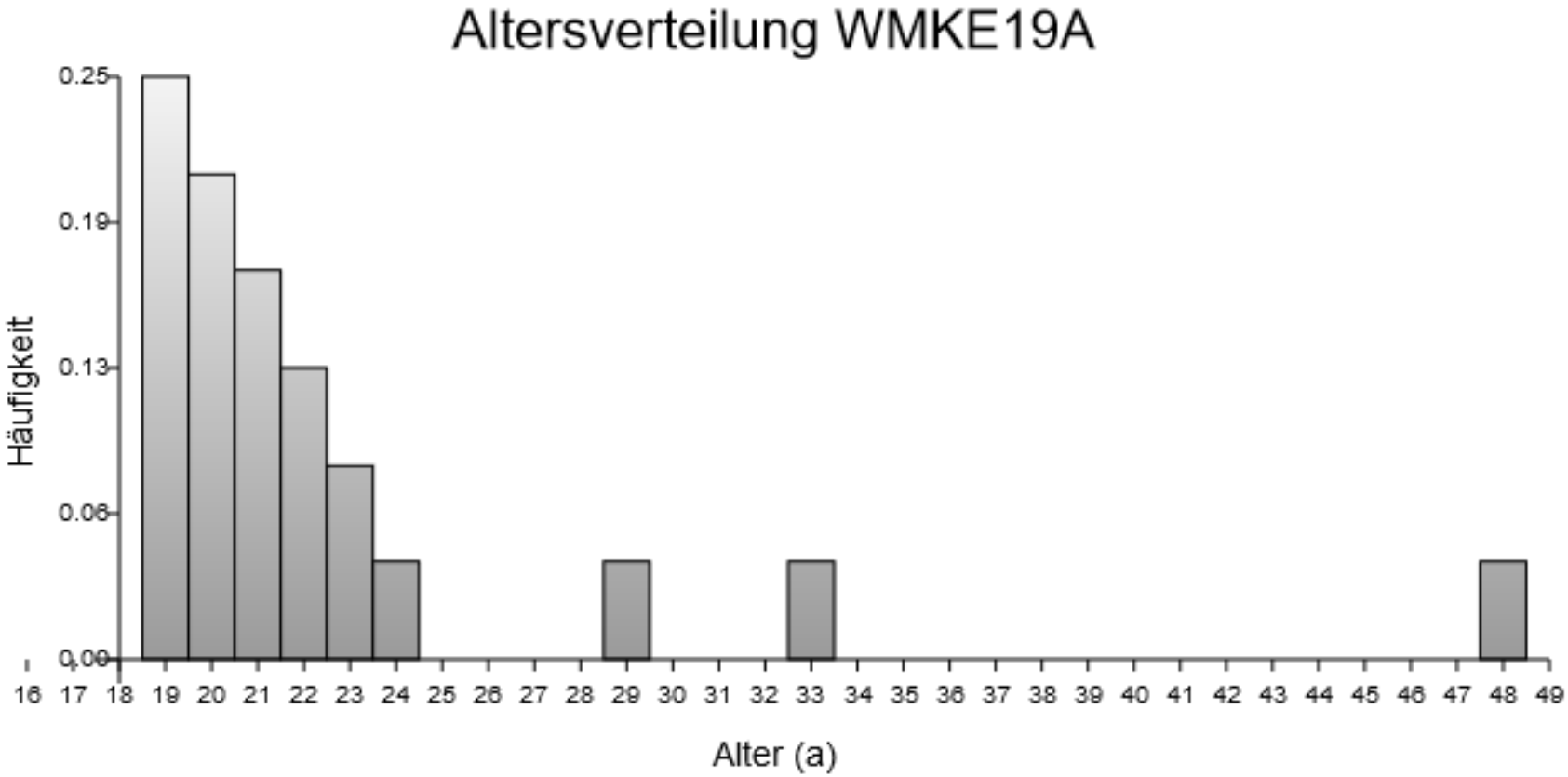
Stichprobe

Die Werte für das Alter lassen sich in eine Häufigkeitsverteilung umschreiben

Alter x_i	Häufigkeit h_i	rel. Häufigkeit f_i	kumuliert F_i
19	6	0.250	0.250
20	5	0.208	0.458
21	4	0.167	0.625
22	3	0.125	0.750
23	2	0.083	0.833
24	1	0.042	0.875
29	1	0.042	0.917
33	1	0.042	0.959
48	1	0.042	1

Stichprobe

Grafik der Häufigkeitsverteilung:



Stichprobe: Lageparameter

Die drei einfachen Lageparameter für die Altersverteilung sind (in a):

- der Modus

$$\bar{x}_M = 19$$

- das arithmetische Mittel

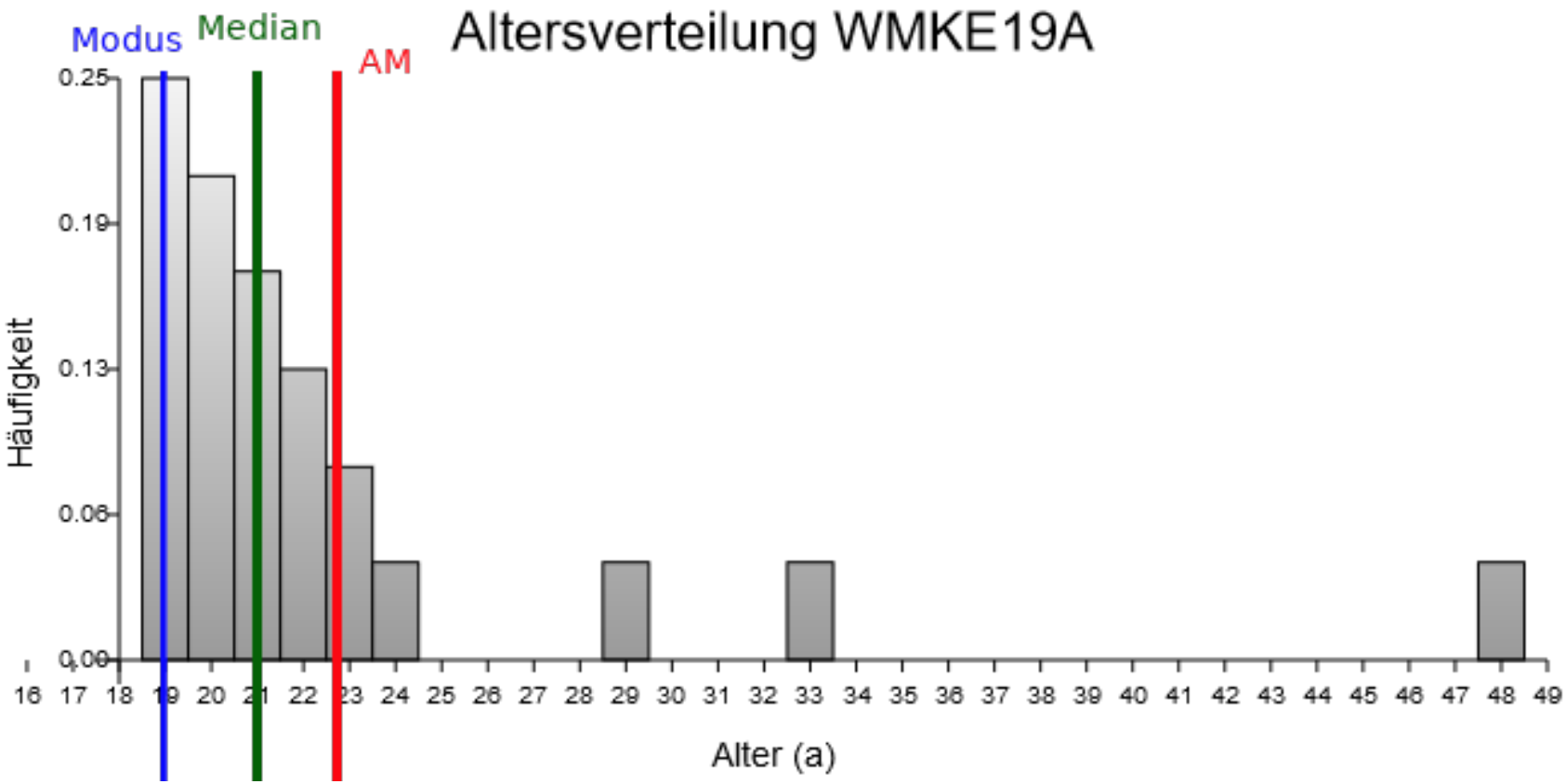
$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{24}(6 \cdot 19 + 5 \cdot 20 + 4 \cdot 21 + 3 \cdot 22 + 2 \cdot 23 + 24 + 29 + 33 + 48) = 22,7$$

- der Median (n ist gerade)

$$\bar{x}_Z = \frac{x_i + x_{i+1}}{2} \text{ mit } i = \frac{n}{2} = 12$$

$$\bar{x}_Z = \frac{x_{12} + x_{13}}{2} = \frac{21 + 21}{2} = 21$$

Stichprobe: Lageparameter



Stichprobe Haarfarbe

Die Verteilung der Haarfarbe im Kurs ist

Farbe x_i	Häufigkeit h_i
blond	9
braun	9
dunkelblond	4
blondbraun	1
straßenköterblond	1

Tabelle: Tabelle

Mit dem nominalen Merkmal ist nicht viel anzufangen, es kann aber der Modus bestimmt werden. Die Ausprägungen 'blond' und 'braun' treten jeweils mit der Häufigkeit $h = 9$ auf, es gibt zwei Werte für den Modus $x_M = \{blond, braun\}$.

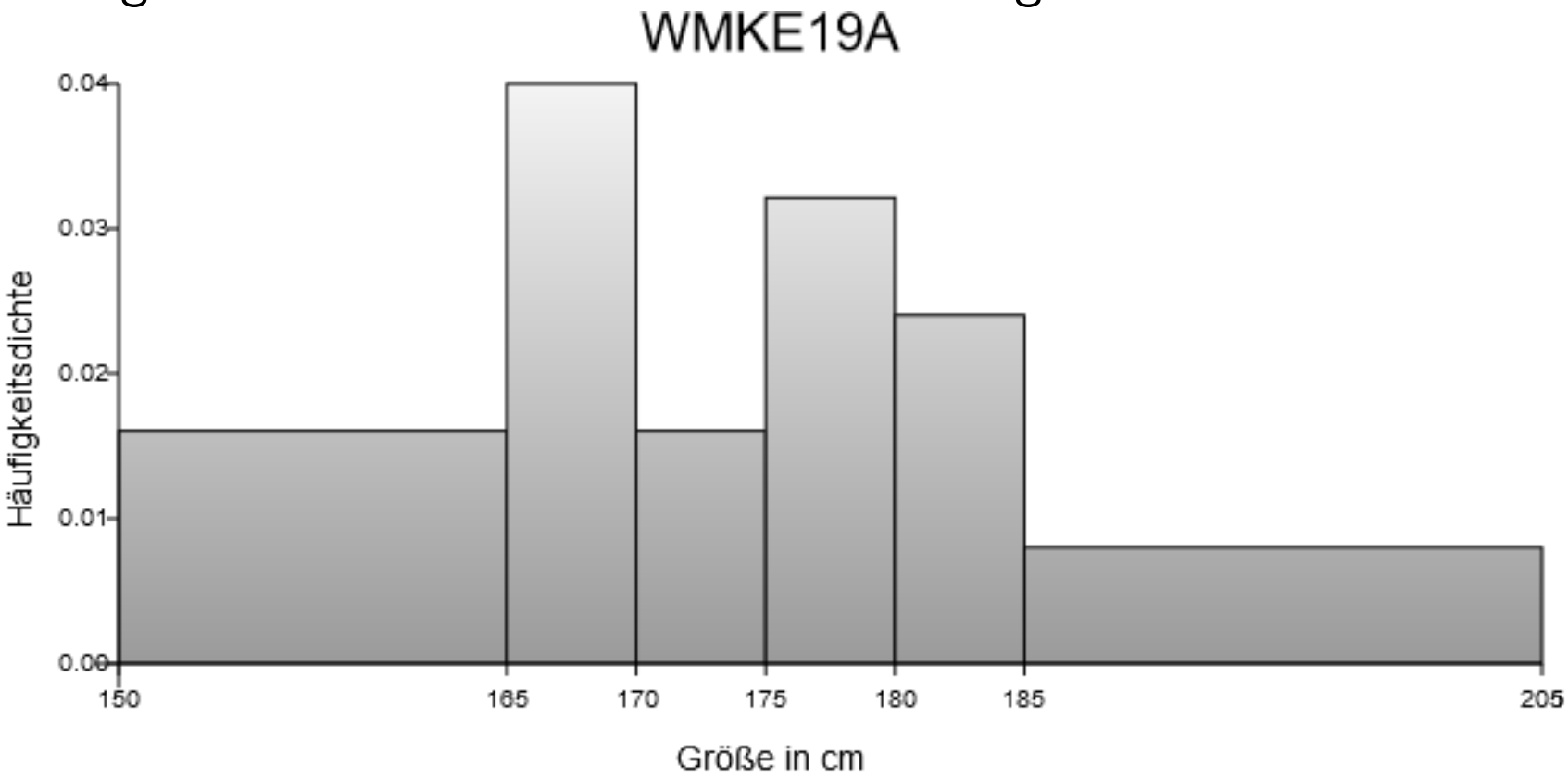
Stichprobe: Größe

Die (bereits klassierten) Daten zur Größenverteilung:
Die Klasse mit der Nummer k entspricht dem Intervall $[x_k^u; x_k^o[$.

k	x_k^u	x_k^o	h_k	f_k	Δ_k	f_k^*
1	150	165	6	0.25	15	0.017
2	165	170	5	0.208	5	0.042
3	170	175	2	0.083	5	0.017
4	175	180	4	0.167	5	0.033
5	180	185	3	0.125	5	0.025
6	185	205	4	0.167	20	0.008

Stichprobe: Größe

Histogramm der klassierten Größenverteilung:



Die *Flächen* der Balken entsprechen den relativen Häufigkeiten f_k. Die Modale Klasse ist die Klasse 2 von 165 bis 170cm.

AM bei klassierten Daten

Im Gegensatz zur vorherigen Tabelle der Verteilung des Lebensalters ist hier bereits Information vernichtet worden. Die Einzelwerte x_i sind nicht mehr bekannt.
Unter der Annahme, dass sich die Werte gleichmäßig in den einzelnen Klassen verteilen, kann mit den Klassenmitten als Näherung für die Berechnung des arithmetischen Mittels gerechnet werden. Die Mitte der Klasse k ist

$$\frac{x_k^u + x_k^o}{2},$$

für das arithmetische Mittel gilt

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \approx \frac{1}{n}\sum_{i=1}^{k} h_k \frac{x_k^u + x_k^o}{2}$$

Beispiel: arithmetisches Mittel bei klassierten Daten.

Für die klassierten Daten (Verteilung der Körpergröße) des Beispiels ergibt sich bei Verwendung der jeweiligen Klassenmitten eine mittlere Körpergröße der 24 Personen im Kurs von

$$\bar{x} = \frac{1}{24}(6 \cdot 157,5 + 5 \cdot 167,5 + 2 \cdot 172,5 + 4 \cdot 177,5 + 3 \cdot 182,5 + 4 \cdot 195) = \frac{4165}{24} = 173,5$$

(alle Angaben in cm). Berechnet man das arithmetische Mittel direkt d.h. ohne Einteilung der Daten in Klassen, so ergibt sich ein Wert von $173,9$ cm. Das arithmetische Mittel der klassierten Daten ist immer eine Näherung.

modale Klasse

Auch der Modus kann bei klassierten Daten nicht direkt angegeben werden. Man rettet sich, indem die Klasse mit der größten auftretenden Häufigkeit als *modale Klasse* festgelegt wird.

Definition: Definition der modalen Klasse

Der Modus einer Stichprobe ist die am häufigsten auftretende Merkmalsausprägung. Liegen statt Einzelwerten klassierte Daten eines Merkmals vor, wird die Klasse mit der größten Häufigkeit modale Klasse genannt.

Beispiel: Größenverteilung im Kurs

Die modale Klasse der Größenverteilung ist die Klasse Nummer 2 von $x_2^u = 165$ cm bis $x_k^o = 170$ cm

Kumulierte Häufigkeitsverteilungen

Eine Stichprobe des Umfangs n enthalte $m \leq n$ unterschiedliche, geordnete Ausprägungen eines ordinalen oder kardinalen Merkmals X, die mit den relativen Häufigkeiten $(f_1, f_2, \ldots f_m)$ auftreten. Unter der **kumulierten absoluten bzw. relativen Häufigkeit H_i bzw. F_i** versteht man die Summe der absoluten oder relativen Häufigkeiten für die Merkmalsausprägungen, die kleiner oder gleich dem jeweiligen Wert x_i sind.

$$H_i = \sum_{j=1}^{i} h_j \qquad \text{bzw.} \qquad F_i = \sum_{j=1}^{i} f_j$$

$$F_1 = f_1 \; ; F_2 = f_1 + f_2 \; ; \ldots$$

Die Vektoren $(H_1, H_2, \ldots H_n)$ bzw $(F_1, F_2, \ldots F_n)$ bilden die kumulierte absolute bzw. relative Häufigkeitsverteilung für die Stichprobe.

Kumulierte Häufigkeitsverteilungen

Beispiel: kumulierte Häufigkeiten für die Altersverteilung.

Die kumulierten absoluten und relativen Häufigkeiten der Lebensalter der Kursteilnehmer sind in der Tabelle aufgeführt:

Alter x_i	Häufigkeit h_i	rel. Häufigkeit f_i	kumuliert F_i
19	6	0.250	0.250
20	5	0.208	0.458
21	4	0.167	0.625
22	3	0.125	0.750
23	2	0.083	0.833
24	1	0.042	0.875
29	1	0.042	0.917
33	1	0.042	0.959
48	1	0.042	1

Verteilungspolygon

Liegt eine Stichprobe eines kardinalen Merkmals X in Form von klassierten Daten mit l Klassen vor, sind die kumulierten relativen Häufigkeiten F_k die Summen der relativen Häufigkeiten für die Klassen 1 bis k.

- Die kumulierte relative Häufigkeit F_k wird wird an der oberen Grenze x_k^o der k-ten Klasse erreicht.
- Die Summe der relativen Häufigkeiten für die Klassen 1 bis l muss Eins ergeben: $F_l = 1$.

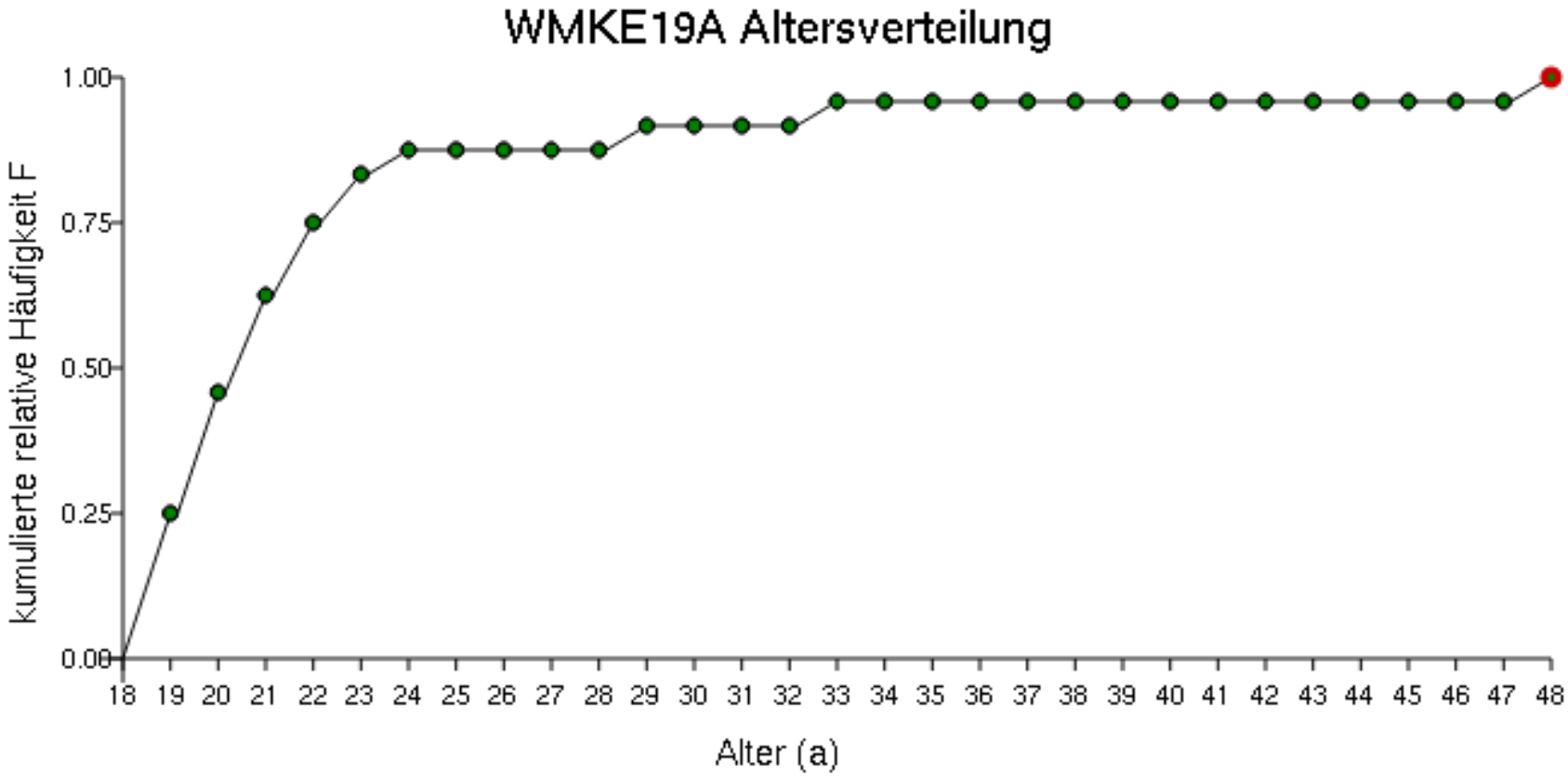
Die Punkte

$$(x_1^o, F_1); (x_2^o, F_2); ...; (x_l^o, F_l)x$$

stellen die Eckpunkte des sogenannten Verteilungspolygons dar. Zusätzlich kann der Startpunkt Punkt $(x_1^u, 0)$ festgelegt werden.

Verteilungspolygon

Trägt man die Punkte in ein Achsenkreuz ein und verbindet die Eckpunkte durch Geraden, so erhält man eine Darstellung des Anteils der Stichprobe, der bis zu einem gegebenen Wert x erreicht ist.

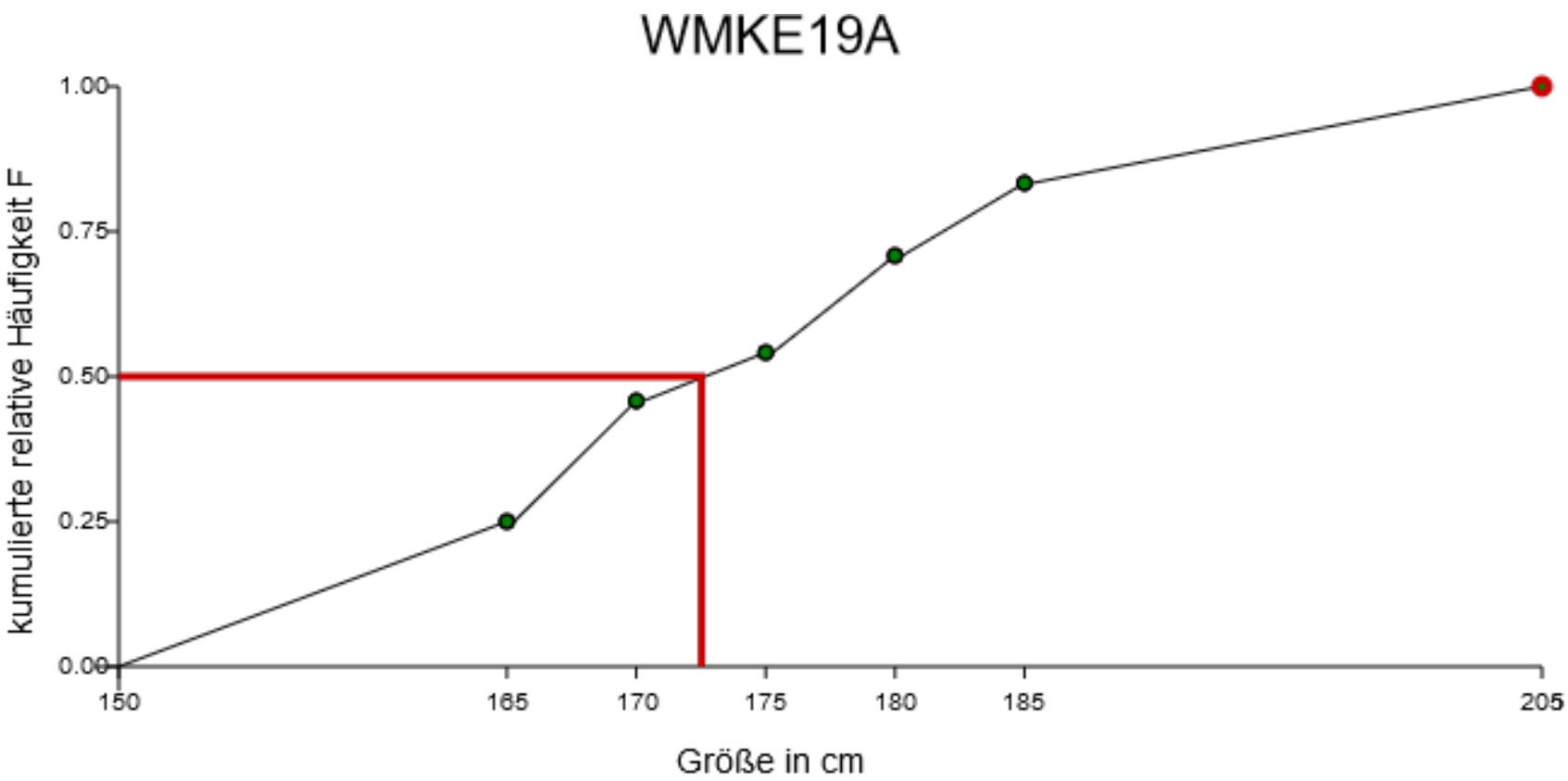


Verteilungspolygon

Interessanter ist das Polygon für die Verteilung der Größen im Kurs: benötigt werden die kumulierten Häufigkeiten

k	x_k^u	x_k^o	h_k	f_k	F_k
1	150	165	6	0.25	0.25
2	165	170	5	0.208	0.458
3	170	175	2	0.083	0.542
4	175	180	4	0.167	0.708
5	180	185	3	0.125	0.833
6	185	205	4	0.167	1

Verteilungspolygon



Eingezeichnet ist der Wert, an dem die kumulierte Häufigkeit den Wert 0,5 erreicht hat (Mitte der geordneten Stichprobe - der Median)

Median bei klassierten Daten

Der Median lässt sich für eine Stichprobe klassierter Daten aus dem Verteilungspolygon ablesen: er ist der zum Funktionswert $F = 0,5 = 50\%$ gehörige x-Wert.

- Der Wert im Verteilungspolygon ist genähert (Geraden entsprechen der gleichmäßigen Verteilung der Werte in jeder Klasse)
- Das Prinzip lässt sich für eine beliebige (kardinale) Stichprobe verallgemeinern
- Fur den Median ist nur die Klasse relevant, in der $F = 0,5$ liegt

Median bei klassierten Daten

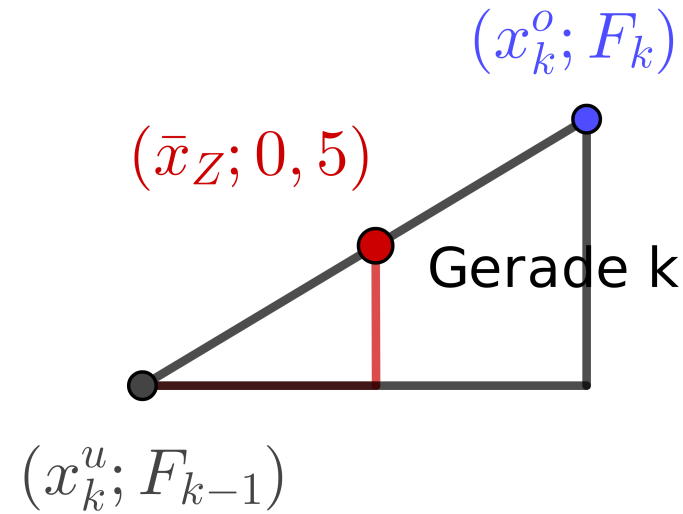
Der Median liegt in der Klasse, in der die kumulierte Häufigkeit über 0,5 steigt (aus F_k ablesbar). Allgemein ist die Gerade in dieser Klasse durch die Punkte $(x_k^u; F_{k-1})$ und $(x_k^o; F_k)$ festgelegt:

Die Steigung der Geraden ist

$$m = \frac{\Delta F}{\Delta x} = \frac{F_k - F_{k-1}}{x_k^o - x_k^u}$$

oder über den Median gerechnet

$$m = \frac{\Delta F}{\Delta x} = \frac{0,5 - F_{k-1}}{\bar{x}_Z - x_k^u}$$



Median bei klassierten Daten

Da die Steigung m in jedem Punkt der Geraden dieselbe ist, gilt

$$\frac{0,5 - F_{k-1}}{\bar{x}_Z - x_k^u} = \frac{F_k - F_{k-1}}{x_k^o - x_k^u}$$

Beispiel: Median der Größenverteilung

Daraus kann mit der Definition des Medians $F(\bar{x}_Z) = 0,5$ die Lage näherungsweise bestimmt werden (in cm):

$$\bar{x}_Z = x_3^u + (x_3^o - x_3^u) \cdot \frac{F(\bar{x}_Z) - F(x_3^u)}{F(x_3^o) - F(x_3^u)}$$

$$= 1,70 + (1,75 - 1,70) \cdot \frac{0,5 - 0,458}{0,542 - 0,458} = 172,5$$

(der exakte Wert ist 171 cm)

Median bei klassierten Daten

Die Verallgemeinerung der Gleichung im Beispiel liefert eine universelle Berechnung des Medians bei klassierten Daten:

Median

- Liegt der Median in der Klasse j, so ist
- der Median genähert

$$\begin{aligned}\bar{x}_Z &= x_j^u + (x_j^o - x_j^u)\frac{F(\bar{x}_Z) - F(x_j^u)}{F(x_j^o) - F(x_j^u)} \\ &= x_j^u + (x_j^o - x_j^u)\frac{F(0,5) - F(x_j^u)}{F(x_j^o) - F(x_j^u)}\end{aligned}$$

Diese Formel resultiert aus der Bedingung $F(\bar{x}_Z) = 0,5$ und der

Quantile

Neben den Mittelwerten sind noch weitere Lageparameter in der Statistik von wesentlicher Bedeutung, die Quantile. Das Quantil ist eine Verallgemeinerung des Median-Konzepts: der Median teilt eine Stichprobe bei 50%

Von Interesse sind daneben beispielsweise auch das 25%-Quantil oder das 75%-Quantil sein (unteres und oberes Quartil). Beim Vergleich von Einkommensverteilungen verschiedener Länder benutzt man Dezentile - das sind die 10%- und 90%-Quantile.

Quantile

Definition: Beschreibende Definition des p-Quantils

für kardinale Merkmale: das $P\%$-Quantil bzw. p-Quantil $\bar{x}_p$ ist der Merkmalswert des Merkmals X, den mindestens $P\%$ aller Merkmalswerte einer Stichprobe vom Umfang n unterschreiten oder höchstenfalls erreichen und den mindestens $(100 - P)\%$ aller Merkmalswerte überschreiten oder zumindest erreichen. Dabei ist $0 < P < 100, p = P/100$ und $0 < p < 1$.

Wie beim Median ist diese Definition nicht eindeutig.

Quantile

Beispiel: Intuitive Berechnung des unteren Quartils.

Bei $n = 11$ geordneten Einzelwerten $(x_1, x_2, ..., x_11)$, ist der 3. Wert das 25% -Quantil (auch als Quartil bezeichnet). Der Wert x_3 erfüllt die beiden notwendigen Bedingungen: es ist der erste Wert, der die Stichprobe in zwei Teilmengen aufteilt - eine Teilmenge mit einem Umfang von mindestens einem Viertel aller Werte besitzt x_1, x_2, x_3 und einer Teilmenge mit mindestens einem Umfang von $3/4$ der Werte $x_3, x_4, ..., x_11$. Im konkreten Beispiel der bereits geordneten Stichprobe

$$(0, 1, 1, 2, 2, 4, 4, 6, 9, 13, 17)$$

ist das untere Quartil $\bar{x}_{0,25} = 1$. Der Wert 1 teilt die Stichprobe in $0, 1, 1$ und $1, 1, 2, 2, 4, 4, 6, 9, 13, 17$.

Quantile

Beispiel: Intuitive Berechnung des unteren Quartils.

Division der Zahl von 11 Elementen durch 4 gibt einen Hinweis auf die Aufteilung der Teilmengen. Die kleinere Teilmenge muss mindestens 3 Elemente - die nächste ganze Zahl nach $2,75$ enthalten, die andere mindestens 9 Elemente - die nächste ganze Zahl nach $8,25 = 3/4 \cdot 11$.

12 geordnete Einzelwerte, z.B. $(0, 1, 1, 2, 2, 3, 4, 4, 6, 9, 13, 17)$, lassen sich einfacher in zwei Teilmengen mit $1/4$ bzw. $3/4$ Umfang aufteilen: $1/4 \cdot 12 = 3$ und $3/4 \cdot 12 = 9$ Elemente. Der 3. und der 4. Wert der Reihe erfüllen die Bedingungen an das untere Quartil. Wie beim Median kann über die Werte gemittelt werden:

$$\bar{x}_{0,25} = \frac{x_3 + x_4}{2} = \frac{1+2}{2} = 1,5$$

Der Wert $1,5$ teilt die Menge von 12 Zahlen in zwei Teilmengen des Umfangs $1/4$ bzw. $3/4$ der Merkmalswerte auf.

Quantile

Definition: Definition des p-Quantils

Das p-Quantil x_p bei n Einzelwerten eines kardinalen Merkmals: Bezeichnet $(x_1, x_2, ..., x_n)$ einen Vektor geordneter, individueller Merkmalswerte eines kardinalen Merkmals X, so wird das p-Quantil x_p in eindeutiger Weise definiert durch

$$\bar{x}_p := \begin{cases} x_i \text{ ,wobei } i = [n \cdot p] + 1, \text{ falls } n \cdot p \text{ nicht ganzzahlig ist,} \\ \frac{x_i + x_{i-1}}{2} \text{ ,wobei } i = [n \cdot p], \text{ falls } n \cdot p \text{ ganzzahlig ist.} \end{cases} \tag{1}$$

Dabei stellen die eckigen Klammern die sogenannten GAUSS-Klammern dar. $[n \cdot p]$ bezeichnet die größte ganze Zahl, die kleiner oder gleich dem Ausdruck $n \cdot p$ innerhalb der Klammer ist.

Quantile

Spezielle Quantile

- das $0,1/0,9$-Quantil heißt auch Dezentil
- das $0,2/0,8$-Quantil heißt auch Quintil
- das $0,25/0,75$-Quantil heißt auch Quartil

Für den Median x_Z, das 50%-Quantil $\bar{x}_{0,5}$, folgt aus der allgemeinen Definition des p-Quantils (1) mit $p = 0,5 = 1/2$ natürlich wieder die Definition des Medians bei Individualdaten eines kardinalen Merkmals.

Quantile

Beispiel: Berechnung des unteren Quartils nach Definition (1).

Bei Berechnung des unteren Quartils ergibt sich für den Vektor

$$(0, 1, 1, 2, 2, 4, 4, 6, 9, 13, 17)$$

von $n = 11$ geordneten Zahlen ein Wert von $\bar{x}_{0,25} = x_3 = 1$, denn $n \cdot p = 11 \cdot 0,25 = 2,75$. Es gilt der erste Teil der Definition mit $[n \cdot p] = [2,75] = 2$ und dem Index $i = 2 + 1 = 3$ des Quartils.

Für den Vektor

$$(0, 1, 1, 3, 2, 2, 4, 4, 6, 9, 13, 17)$$

von $n = 12$ geordneten Zahlen ergibt sich wie oben auch $\bar{x}_{0,25} = (x_3 + x_4)/2 = (1 + 2)/2 = 1,5$. Wegen $n \cdot p = 12 \cdot 0,25 = 3$ wird in diesem Fall der zweite Teil der Definition benutzt, wobei $[n \cdot p] = [3] = 3$.

Quantile bei klassierten Daten

Bei klassierten Daten kann das p-Quantil $\bar{x}_p$ grafisch mit Hilfe des Verteilungspolygons $F(x)$ bestimmt werden. Das p-Quantil ist als der zum Funktionswert $F(x) = p$ gehörige Variablenwert definiert:

$$F(\bar{x}_p) = p \tag{2}$$

Es kann analog zum Fall des Medians rechnerisch durch eine allgemeine Formel berechnet werden:

$$\begin{aligned}\bar{x}_p &= x_j^u + (x_j^o - x_j^u) \cdot \frac{F(x_p) - F(x_j^u)}{F(x_j^o) - F(x_j^u)} \\ &= x_j^u + (x_j^o - x_j^u) \cdot \frac{p - F(x_j^u)}{F(x_j^o) - F(x_j^u)}\end{aligned}$$

wenn das Quantil in der Klasse j liegt.

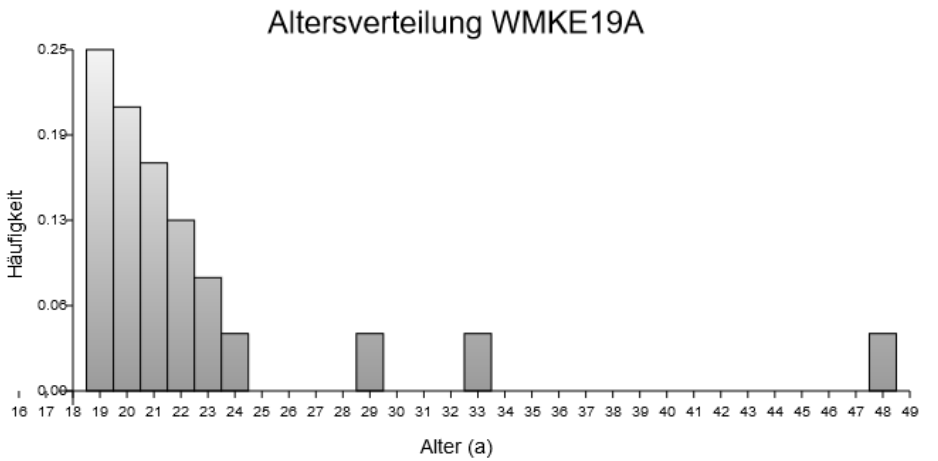
Quantile bei klassierten Daten

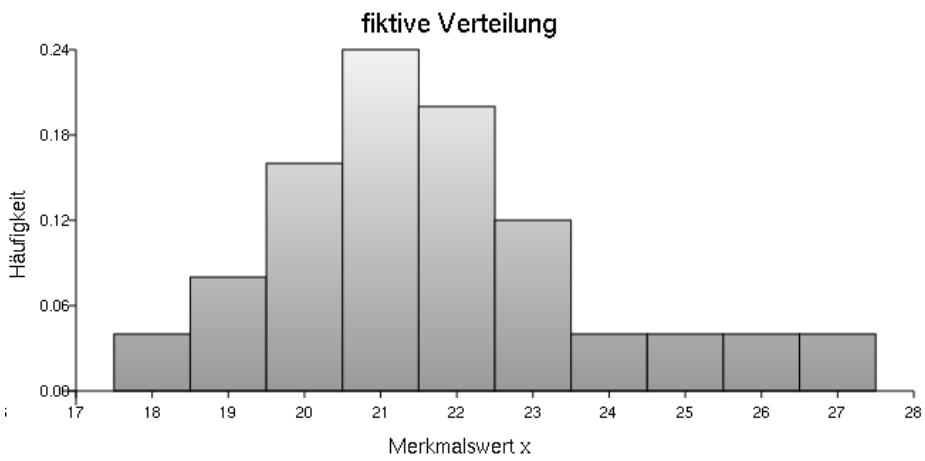
Die Berechnung ergibt sich aus der Bedingung $F(\bar{x}_p) = p$ und der Konstruktion des Verteilungspolygons, bei der die Eckpunkte $(x_j^u; F(x_j^u))$ und $(x_j^o; F(x_j^o))$ durch Geraden verbunden werden. Vor der Bestimmung eines Quantils über die Formel muss wieder - grafisch oder mit Hilfe einer Tabelle - ermittelt werden, in welcher Klasse j das p-Quantil liegt.

Setzt man für $p = 0,5$, so ergibt sich aus natürlich wieder die Formel zur Berechnung des Medians für klassierte Daten.

Streuung

Die beiden Verteilungen unten besitzen dasselbe arithmetische Mittel. Der Lageparameter Mittelwert reicht offensichtlich nicht aus, um die Verteilungen sinnvoll zu beschreiben, bei der Bestimmung geht zuviel der relevanten Information verloren. Die Verteilungen unterscheiden sich deutlich in der typischen Abweichung einzelner Werte vom gemeinsamen arithmetischen Mittel.





Streuung

Lagemaße (Mittelwerte) geben typische Werte einer Stichprobe an. Streumaße erweitern die Idee um die Aussage, ob die einzelnen Merkmalswerte dicht bei einem Mittelwert liegen oder eher davon abweichen. Da Abstände aber nur für kardinale Merkmale sinnvoll sind, können Streumaße zunächst nur für kardinale Merkmale angegeben werden.

Die Spannweite ist das primitivste Streuungsmaß, sie gibt die Differenz zwischen dem größten und dem kleinsten Merkmalswert einer Stichprobe $(x_1, x_2, ...x_n)$ eine kardinalen Merkmals X an

Definition: Spannweite

$$s_W = \max\{x_1, x_2, ...x_n\} - \min\{x_1, x_2, ...x_n\}. \tag{3}$$

Bei klassierten Daten ist das Maximum bzw. Minimum jedoch nicht bekannt, hier wird die Spannweite als Differenz zwischen der oberen Klassengrenze der obersten von k Klassen x_k^o und der untersten Klassengrenze der ersten Klasse x_1^u definiert:

$$s_W := x_k^o - x_1^u \text{ für klassierte Daten in } k \text{ Klassen} \tag{4}$$

Streuung

- Die Aussagekraft der Spannweite für eine Stichprobe $(x_1, x_2, ..., x_n)$ ist sehr eingeschränkt, da s_W nur aus zwei Werten dieser Stichprobe berechnet wird: die einfache Berechnung wird mit einem hohen Informationsverlust erkauft. Ist mindestens einer der Werte ein Ausreißer, hilft die Spammweite wenig weiter.
- Spannweiten verschiedener Stichproben unterschiedlichen Umfangs lassen sich nicht miteinander vergleichen, da bei der Berechnung der Spannweite die Größe des Stichprobenumfangs nicht berücksichtigt wird.

Streuung

Die Streumaße entsprechen Werten für die typische Abweichung der Merkmalswerte von einem Lageparameter. Es liegt nahe, Streuung der Merkmalswerte $(x_1, x_2, ..., x_n)$ eines kardinalen Merkmals X über die Summe der Abweichungen der Einzelwerte x_i von einem Mittelwert (meist das AM) $\Delta_i = (x_i - \bar{x})$ messen zu wollen.

Die Summe der Einzelabweichungen

$$\sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x}$$

$$= n\frac{1}{n}\sum_{i=1}^{n} x_i - n\bar{x} = n\bar{x} - n\bar{x} = 0$$

verschwindet aber, weil die Abweichungen in zwei Richtungen gemessen werden.

Streuung

Um ein Maß für die Summe der Abweichungen vom Mittelwert zu erhalten, kann das Vorzeichen über Absolutbeträge der Abweichungen eliminiert werden.

Definition: mittlere absolute Abweichung

$$d_{\bar{x}} := \frac{1}{n} \sum_{i=1}^{n} |x_i - \bar{x}|. \tag{5}$$

bzw. bei der mittleren absoluten Abweichung vom Median $\bar{x}_Z$:

$$d_{\bar{x}_Z} := \frac{1}{n} \sum_{i=1}^{n} |x_i - \bar{x}_Z|. \tag{6}$$

Oft wird das zweite Maß bevorzugt, weil der Median $\bar{x}_Z$ diese Summe der absoluten Abweichungen minimiert

Streuung

Wegen der Beträge kann die mittlere absolute Aweichung nicht negativ werden. Sie kann nur dann den Wert Null annehmen, wenn alle Einzelwerte identisch sind - damit ist automatisch auch der jeweilige Mittelwert festgelegt:

$$x_1 = x_2 = ... = x_n = \bar{x} \text{ bzw. } \bar{x}_Z$$

verschwinden alle Differenzen in der Berechnung, es gibt natürlich keine Streuung unter den Merkmalswerten, die beiden Maße werden also zu Null

Beispiel: Berechnung der mittleren absoluten Abweichungen

Der Median der Altersverteilung im Kurs hat einen Wert von 21 Jahren
Die mittlere absolute Abweichung der Lebensalter vom Median in Jahren beträgt

$$d_{\bar{x}_Z} = \frac{1}{24}(6 \cdot |19-21| + 5 \cdot |20-21| + 4 \cdot |21-21| + 3 \cdot |21-22| + 2 \cdot |23-21| + |24-21| + |29-21| + |33-21| + |48-21|) = 3,08.$$

und die mittlere absolute Abweichung vom AM mit $\bar{x} = 22,7$ in Jahren

$$d_{\bar{x}} = \frac{1}{24}(6 \cdot |19-22,7| + 5 \cdot |20-22,7| + 4 \cdot |21-22,7| + 3 \cdot |22,7-22| + 2 \cdot |23-22,7| + |24-22,7| + |29-22,7| + |33-22,7| + |48-22,7|) = 3,67.$$

Streuung

Zur Berechnung der mittleren absoluten Abweichungen vom arithmetischen Mittel oder Median werden alle Werte der Stichprobe $(x_1, x_2, ..., x_n)$ verwendet. Bei der Berechnung wird also im Gegensatz zur Spannweite keine Information ignoriert.

Ein zum Vergleich der Streuungen verschiedener Stichproben konzipiertes Streuungsmaß sollte den Umfang der jeweiligen Stichprobe berücksichtigen. Das ist bei den beiden Maßen der Fall.

Das am meisten verwendete Streuungsmaß ist allerdings die empirische Standardabweichung, die ganz wesentlich auf der empirischen Varianz basiert.

empirische Varianz und Standardabweichung

Die zweite Möglichkeit, die Vorzeichen zu eliminieren ist das Quadrat. Die *durchschnittliche quadratische Abweichung* der Einzelwerte einer (kardinalen) Stichprobe $(x_1, x_2, ..., x_n)$ vom arithmetischen Mittel $\bar{x}$ wird als die empirische Varianz s^2 bezeichnet:

Definition: empirische Varianz

$$s^2 := \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{7}$$

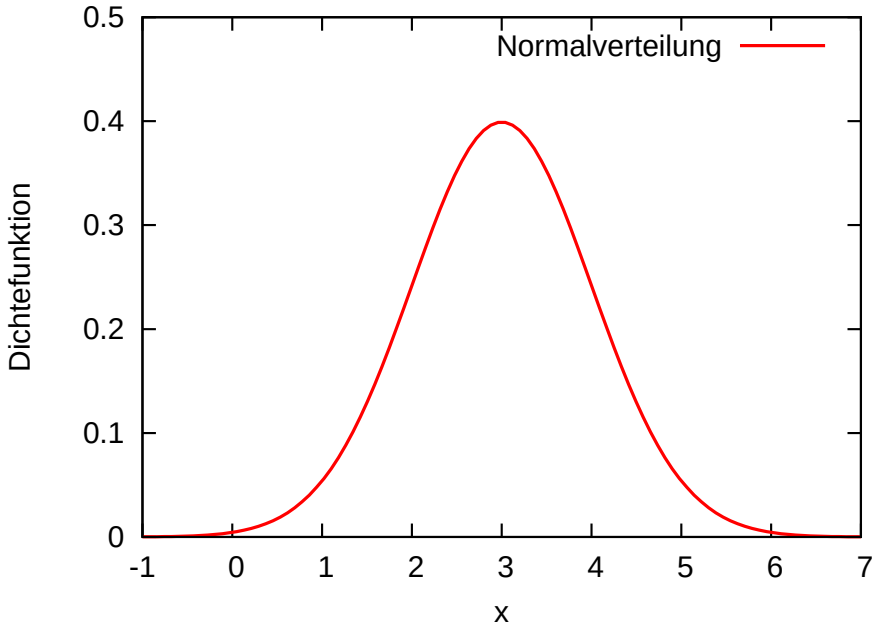
Die Wurzel der Varianz wird als Standardabweichung bezeichnet:

$$s = \sqrt{s^2}$$

Auch die Varianz und die Standardabweichung sind immer positiv.

Varianz: Hintergrund

Die Wahrscheinlichkeitsdichte einer normalverteilten Zustandsgröße wird durch die Gauß-Verteilung (Glockenkurve) beschrieben:



$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Die Funktion enthält zwei Parameter: μ legt die Stelle des Maximums fest, der Wert σ die Breite, das Quadrat σ^2 wird als Varianz der Verteilung bezeichnet.

Varianz: Hintergrund

Geht man davon aus, dass die Häufigkeitsverteilung einer Stichprobe durch zufällige Abweichungen entsteht, werden sie im Idealfall durch eine Gauß-Verteilung bechrieben: für $n \to \infty$ wird die Häufigkeitsverteilung zur Glockenkurve.

Die Häufigkeitsverteilung $f(x)$ kann durch eine Gauß-Kurve abgeschätzt werden. Die bekannten Lageparameter entsprchen dann einer Näherung für den Maximalwert μ der Kurve. Die mittlere quadratische Abweichung vom AM ist eine geeignete Näherung für die Varianz σ^2.

Die Gauß-Verteilung ist mit den beiden Parametern für die Breite und die Lage des Maximums vollständig beschrieben, Lageparameter und Streumaße sind Näherungen für die beiden theoretischen Größen.

Beispiel: Streuung der Altersverteilung

Für die Alterverteilung im Kurs können jetzt die mittlere absolute Abweichung vom AM $d_{\bar{x}}$ und die Varianz berechnet werden:

x_i	h_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$\lvert x_i - \bar{x} \rvert$
19	6	-3,67	13,44	3,67
20	5	-2,67	7,11	2,67
21	4	-1,67	2,78	1,67
22	3	-0,67	0,44	0,67
23	2	0,33	0,11	0,33
24	1	1,33	1,78	1,33
29	1	6,33	40,11	6,33
33	1	10,33	106,78	10,33
48	1	25,33	641,78	25,33

Tabelle: Tabelle

Beispiel: Streuung der Altersverteilung

die mittlere absolute Abweichung vom AM ist (in a)

$$d_{\bar{x}} = \frac{1}{n}\sum_{i=1}^{m} h_i|x_i - \bar{x}| = \frac{1}{n}(6 \cdot |19 - 22,7| + 5 \cdot |20 - 22,7| + 4 \cdot |21 - 22,7|$$
$$+ 3 \cdot |22 - 22,7| + 2 \cdot |23 - 22,7| + |24 - 22,7| + |29 - 22,7| + |33 - 22,$$
$$= 3,67$$

die Varianz (in a^2)

$$s^2 = \frac{1}{n}\sum_{i=1}^{m} h_i|x_i - \bar{x}| = \frac{1}{n}(6 \cdot (19 - 22,7)^2 + 5 \cdot (20 - 22,7)^2 + 4 \cdot (21 - 22,$$
$$+ 3 \cdot (22 - 22,7)^2 + 2 \cdot (23 - 22,7)^2 + (24 - 22,7)^2 + (29 - 22,7)^2 + (33$$
$$= 38,31$$

Streuung

Das Beispiel zeigt, dass die Varianz zwar gut begründet, aber nicht besonder sinnvoll ist (Angabe in Quadratjahren), das bessere Streumaß ist die Standarabweichung

Beispiel: Standardabweichung der Altersverteilung

Die (empirische) Standardabweiichung ist die Wurzel der Varianz:

$$s := \sqrt{s^2}$$

für die Altersverteilung im Kurs ist

$$s = \sqrt{38,31\ \text{a}^2} = 6,19\ \text{a}$$