

WIRTSCHAFTSSTATISTIK

MODUL 5: STREUUNGSPARAMETER

WS 2020/21

DR. E. MERINS

EINLEITUNG

Problem der Lageparameter:

Die Lageparameter schweigen sich aus über die Streuung der Daten. Das arithmetische Mittel (der Durchschnitt) und auch der Median verdecken oft eine große Ungleichheit.

Statistiker-Witz (frei nach Franz Josef Strauß):

Zwei Männer sitzen im Wirtshaus.

Der eine verdrückt eine ganze Kalbshaxe, der andere trinkt zwei Maß Bier.

Statistisch (im Mittelwert) gesehen ist das für jeden eine Maß Bier und eine halbe Haxe.

Aber in Wirklichkeit der eine hat sich überfressen, und der andere ist besoffen.

→ die Berechnung des Durchschnitts ist nicht immer sinnvoll

→ der Durchschnitt kann offensichtlich nicht immer alles beschreiben

STREUUNG UM DEN MITTELWERT

Beispiel:

In der folgenden Häufigkeitstabelle und den darauf folgenden Säulendiagrammen ist die Notenverteilung zweier Schülergruppen (Mädchen und Jungen) dargestellt, deren Mittelwert gleich ist.

Schüler Nr.	1	2	3	4	5	6	7	8	9	10	
Note Mädchen	3,2	3,5	2,9	3,3	3,4	2,5	2,7	2,8	3,1	2,6	$\bar{x}=3,0$
Note Jungs	1,0	1,0	2,0	2,5	3,2	2,8	3,5	2,0	6,0	6,0	$\bar{x}=3,0$

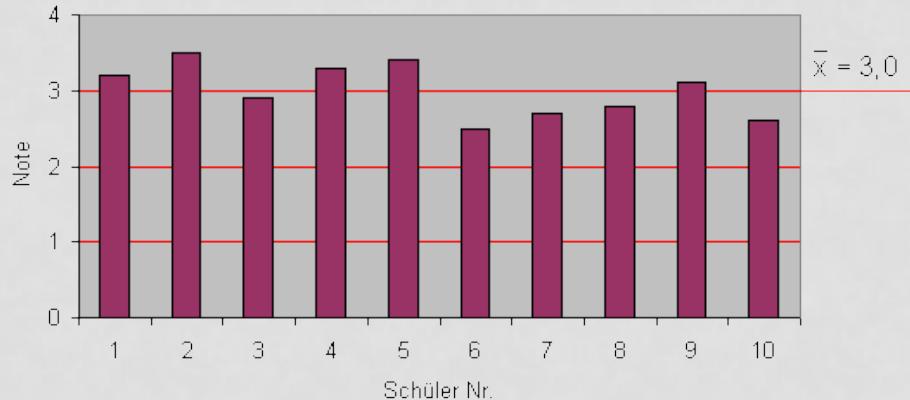
$$\bar{x}_{\text{Mädchen}} = 1/10 * (3,2 + 3,5 + 2,9 + 3,3 + 3,4 + 2,5 + 2,7 + 2,8 + 3,1 + 2,6) = 3,0$$

$$\bar{x}_{\text{Jungs}} = 1/10 * (1,0 + 1,0 + 2,0 + 2,5 + 3,2 + 2,8 + 3,5 + 2,0 + 6,0 + 6,0) = 3,0$$

STREUUNG UM DEN MITTELWERT

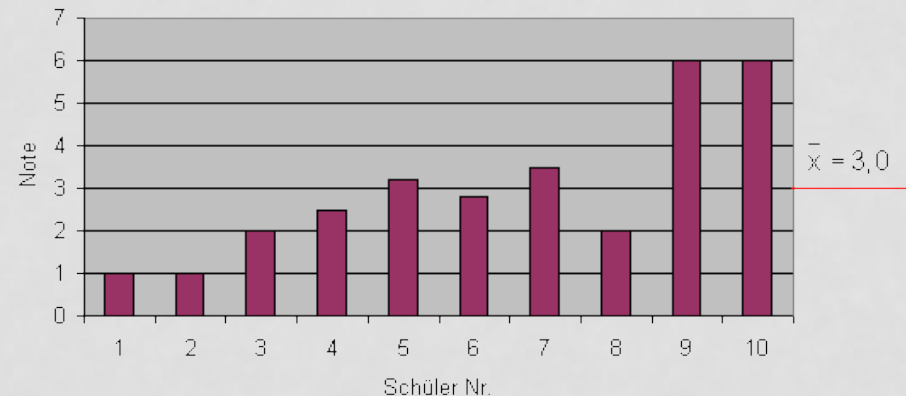
Beispiel:

Notenverteilung Mädchen:



Die Noten liegen alle sehr nahe am Mittelwert
→ Sie streuen wenig um den Mittelwert

Notenverteilung Jungen:



Die Abweichungen vom Mittelwert sind groß
→ Sie streuen stark um den Mittelwert

Die Statistik bietet Möglichkeiten, die **Streuung** näher zu untersuchen und mit Hilfe der **Streuungsparametern** die Streuung zu beschreiben.

STREUUNGSPARAMETER

Forderungen an eine „gute“ Kennzahl zur Messung der Streuung:

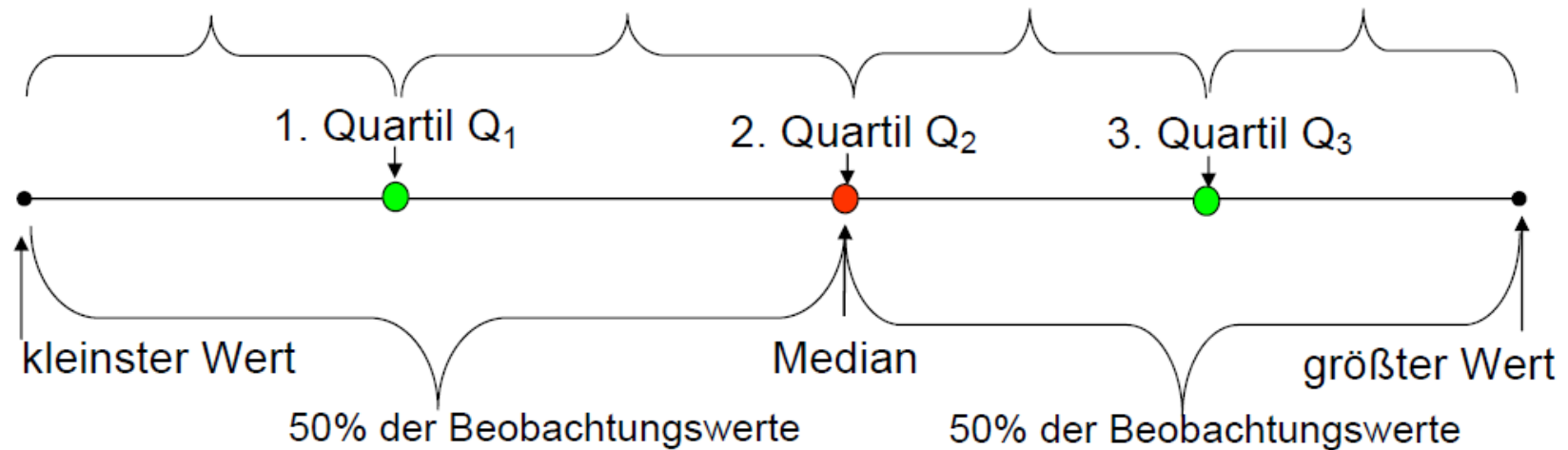
- Bezugspunkt, um den die Werte streuen (→ Lageparameter)
- alle Beobachtungswerte werden berücksichtigt
- Streuung = 0 (alle Werte sind gleich) → Streuungsparameter = 0
- je größer die Streuung, umso größer der Streuungsparameter
- der Streuungsparameter ist unabhängig von der Anzahl der Beobachtungswerte n

QUARTIL

5-Punkte-Zusammenfassung der geordneten statistischen Reihe:

25% der Beobachtungswerte

25% der Beobachtungswerte



Der (Inter-)**Quartilsabstand** (engl.: *interquartile range*, **IQR**) bezeichnet die Differenz zwischen dem oberen und dem unteren Quartil $Q_3 - Q_1$ und umfasst daher 50% der Verteilung.

Der Quartilsabstand wird als **Streuungsmaß** verwendet.

QUARTILSABSTAND

Zusammenfassung:

Der Median teilt einen nach Größe sortierten Datensatz in der Mitte

→ links und rechts vom Median liegen gleich viele Beobachtungswerte. Unterteilt man die linke und die rechte Hälfte nach gleicher Vorschrift, wie man den Median bestimmt, so erhält man 4 gleich große Bereiche, die durch drei Quartils aufgeteilt werden.

25% aller geordneten Beobachtungswerte sind kleiner als das 1. Quartil.

50% aller geordneten Beobachtungswerte sind kleiner als das 2. Quartil.

75% aller geordneten Beobachtungswerte sind kleiner als das 3. Quartil.

Zwischen dem 1. und 3. Quartil liegen 50% aller Beobachtungswerte.

Dieser Bereich wird auch **Quartilsabstand** genannt.

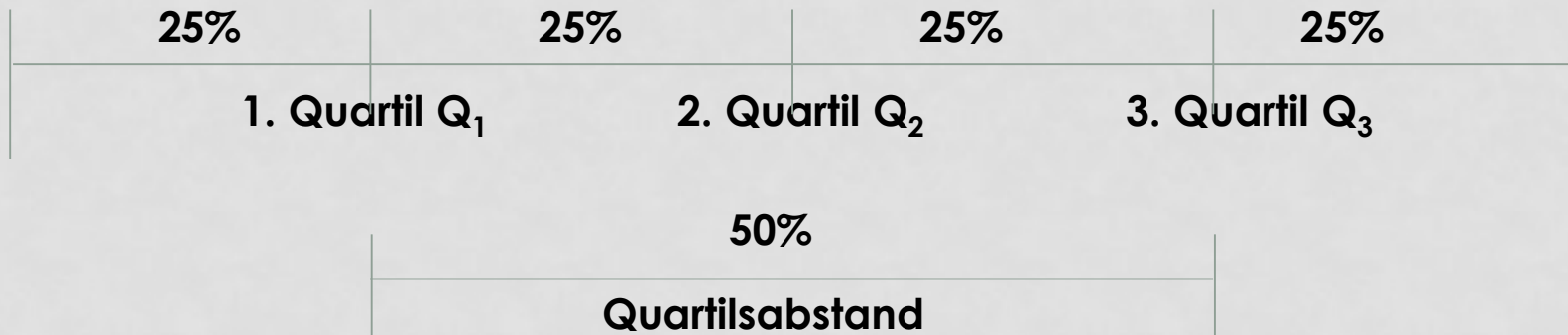
QUARTILSABSTAND

Beispiel:

Die Liste enthält von 13 Schülern die Körpergröße.

Die Merkmalsausprägungen (Beobachtungswerte) wurden nach der Größe geordnet.

Schüler Nr.	1	2	3	4	5	6	7	8	9	10	11	12	13
Größe in m	1,60	1,67	1,67	1,68	1,68	1,70	1,70	1,72	1,73	1,75	1,76	1,78	1,84



QUARTILSABSTAND

Beispiel:

Schüler Nr.	1	2	3	4	5	6	7	8	9	10	11	12	13
Größe in m	1,60	1,67	1,67	1,68	1,68	1,70	1,70	1,72	1,73	1,75	1,76	1,78	1,84
	25%			25%				25%				25%	
	1. Quartil Q_1			2. Quartil Q_2				3. Quartil Q_3					
				50%									
	Quartilsabstand												

$$\bar{x}_Z = Q_2 = x_{\frac{n+1}{2}} = x_{\frac{13+1}{2}} = x_7 = 1,70$$

$$1. \text{ Quartil: } Q_1 = \frac{1}{2}(x_3 + x_4) = \frac{1}{2}(1,67 + 1,68) = 1,675$$

$$3. \text{ Quartil: } Q_3 = \frac{1}{2}(x_{10} + x_{11}) = \frac{1}{2}(1,75 + 1,76) = 1,755$$

$$Q_A = IQR = Q_3 - Q_1 = 1,755 - 1,675 = 0,08$$

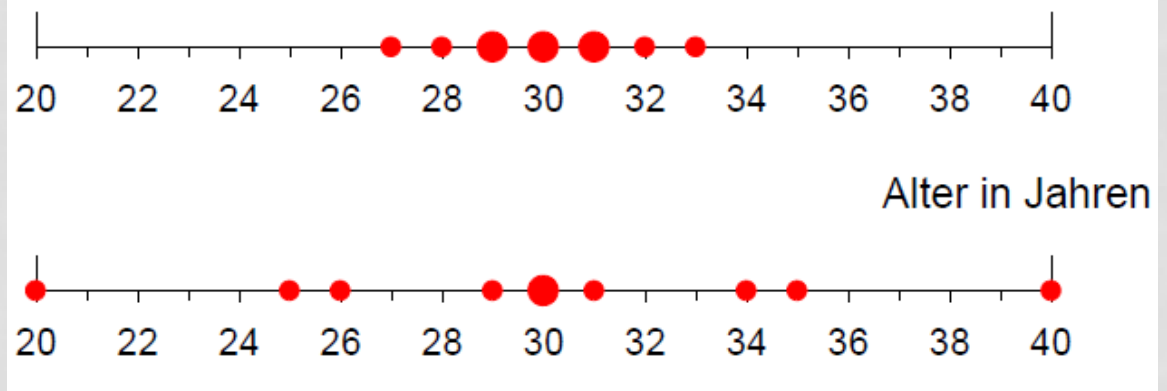
SPANNWEITE

Spannweite (oder Variationsbreite) w : Ausdehnung der Werte (Maß für die Breite des Streubereichs einer Häufigkeitsverteilung)

Für ordinale und metrische Merkmale gilt:

$$w = x_{\max} - x_{\min}$$

Fall 1: $w = 33 - 27 = 6$



Fall 2: $w = 40 - 20 = 20$

SPANNWEITE

$$W = x_{\max} - x_{\min}$$

Beispiel:

Schüler Nr.	1	2	3	4	5	6	7	8	9	10
Note Mädchen	3,2	3,5	2,9	3,3	3,4	2,5	2,7	2,8	3,1	2,6
Note Jungs	1,0	1,0	2,0	2,5	3,2	2,8	3,5	2,0	6,0	6,0

$$w_{\text{Mädchen}} = 3,5 - 2,5 = 1$$

$$w_{\text{Jungs}} = 6,0 - 1,0 = 5,0$$

QUARTILSABSTAND VS. SPANNWEITE

Vergleich zwischen Quartilsabstand und Spannweite:

Quartilsabstand

Von Ausreißern unabhängig

Gibt die Breite des mittleren Bereichs an, in dem ca. 50% aller Werte liegen

Spannweite

Vom kleinsten und größten Wert abhängig

Gibt die Gesamtbreite an, in dem alle Werte liegen

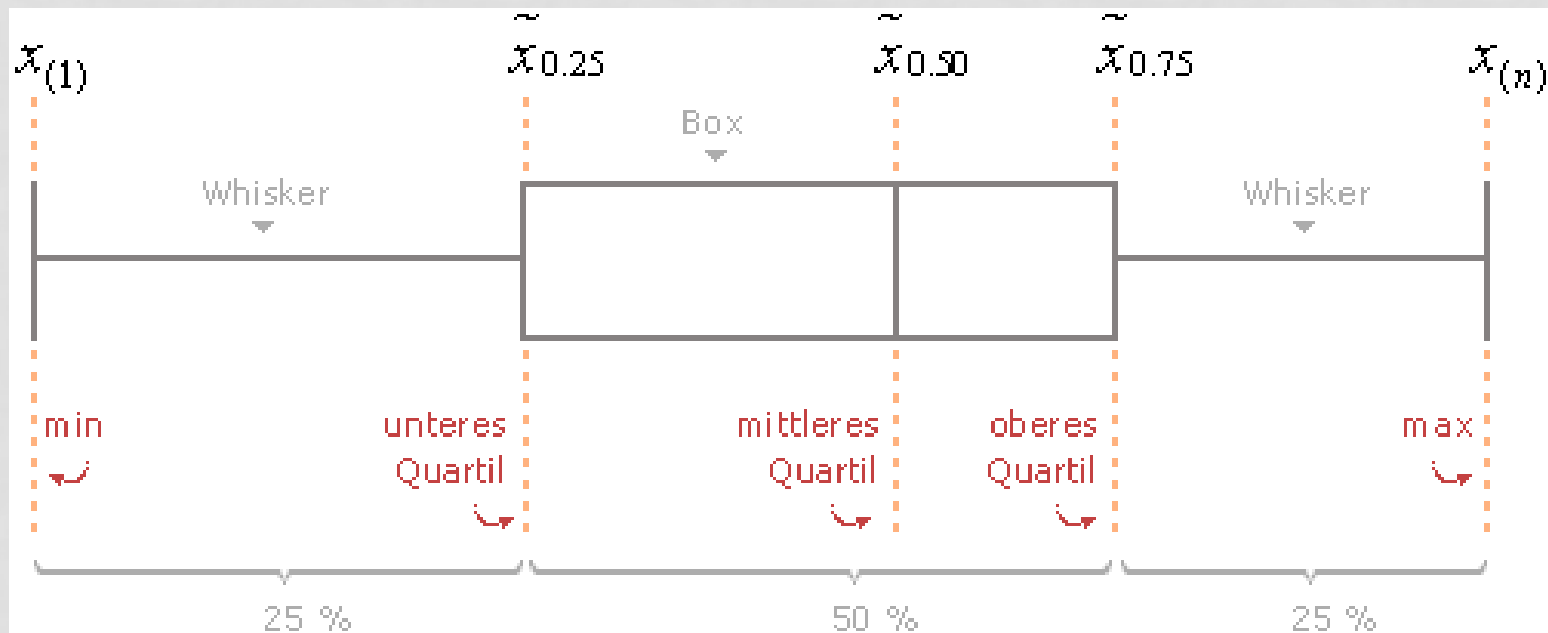
BOXPLOT

Die grafische Darstellung der 5-Punkte-Zusammenfassung heißt

Box-and-Whisker-Plot

Die 5-Punkte-Zusammenfassung besteht aus:

Minimum, Q1, Median, Q3, Maximum



BOXPLOT

Aus einem **Boxplot** lassen sich Informationen über die:

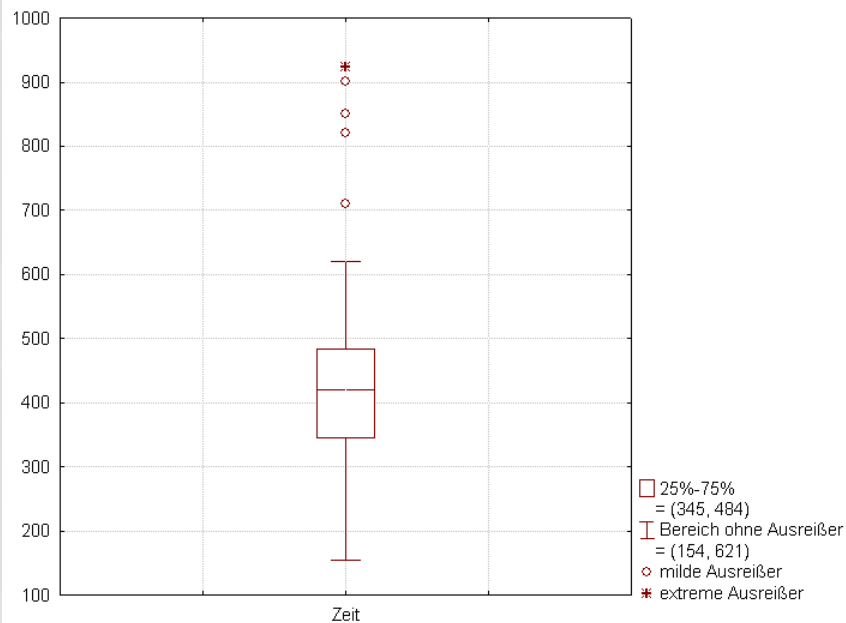
- Lokalisation (Lage des Median)
- Streuungsmaße:
 - **Spannweite** → Ausdehnung eines Boxplots (Differenz $w = x_{\max} - x_{\min}$)
 - **Quartilsabstand** → Ausdehnung der Box (Differenz $IQR = Q_3 - Q_1$)
- Schiefe (Vergleich der beiden Hälften der Box oder der Längen der Whisker)

eines Datensatzes sowie über den evtl. vorliegenden Ausreißer gewinnen.

Eine der Definitionen der Whisker besteht darin, die Länge der Whisker auf maximal das 1,5-Fache des **Interquartilsabstands** ($1,5 \times IQR$) zu beschränken. Der Whisker endet nicht genau nach dieser Länge, sondern bei dem Wert aus den Daten, der noch innerhalb dieser Grenze liegt. Die Länge der Whisker wird also durch die Datenwerte und nicht allein durch den IQR bestimmt. Dies ist auch der Grund, warum die Whisker nicht auf beiden Seiten gleich lang sein müssen. Gibt es keine Werte außerhalb der Grenze von $1,5 \times IQR$, wird die Länge des Whiskers durch den maximalen und minimalen Wert festgelegt. Andernfalls werden die Werte außerhalb der Whisker separat in das Diagramm eingetragen.

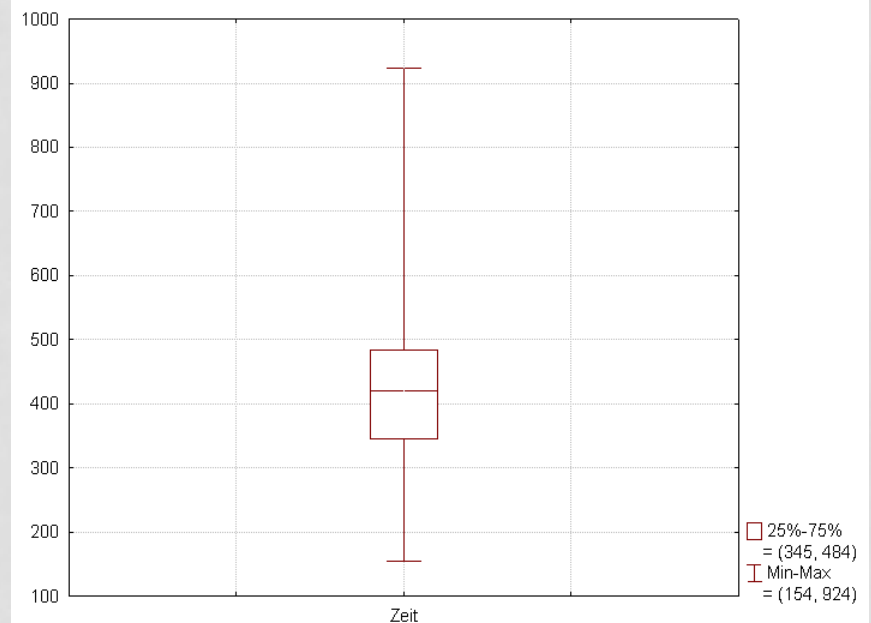
BOXPLOT

Beispiel:



Quartilsabstand: $484 - 345 = 139$

Spannweite: $621 - 154 = 467$



Quartilsabstand: $484 - 345 = 139$

Spannweite: $924 - 154 = 770$

Häufig werden Ausreißer, die zwischen $1,5 \times \text{IQR}$ und $3 \times \text{IQR}$ liegen, als „milde“ Ausreißer bezeichnet und Werte, die über $3 \times \text{IQR}$ liegen, als „extreme“ Ausreißer.

VARIANZ

In der beschreibenden Statistik nennt man das arithmetische Mittel der Abweichungsquadrate die **Varianz**.

Eigenschaften:

- wichtiger Streuungsparameter
- Voraussetzung: metrisches Merkmal
- Ausgangswert für weitere folgende Streuungsparameter:
 - **Standardabweichung**
 - **Variationskoeffizient**

→ **Mittelwert und Varianz bzw. Standardabweichung hängen eng zusammen.**

VARIANZ

Konstruktion der Varianz:

Bezugspunkt:

$$\bar{x}$$

Einzelstreuung/Einzelabweichung:

$$(x_i - \bar{x})$$

Summe der Einzelabweichungen:

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

Summe der quadratischen Abweichungen:

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

Varianz:

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

$$s^2 = \frac{1}{n} ((x_1 - \bar{x})^2 + \dots + (x_n - \bar{x})^2) = \underbrace{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}_{\text{Formel (1)}} = \underbrace{\left(\frac{1}{n} \sum_{i=1}^n x_i^2 \right) - \bar{x}^2}_{\text{Formel (2)}}$$

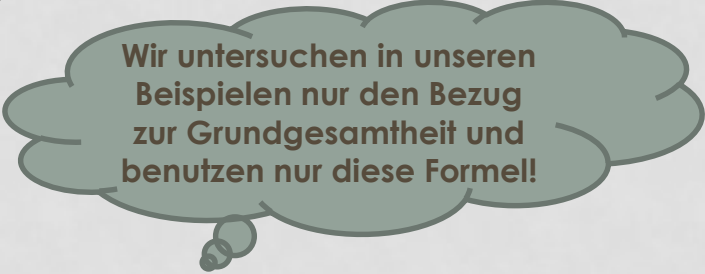
VARIANZ

Konstruktion der Varianz:

Bemerkung:

Handelt es sich bei den zu untersuchenden Daten um die Grundgesamtheit (Population), dann wird mit $1/n$ gewichtet:

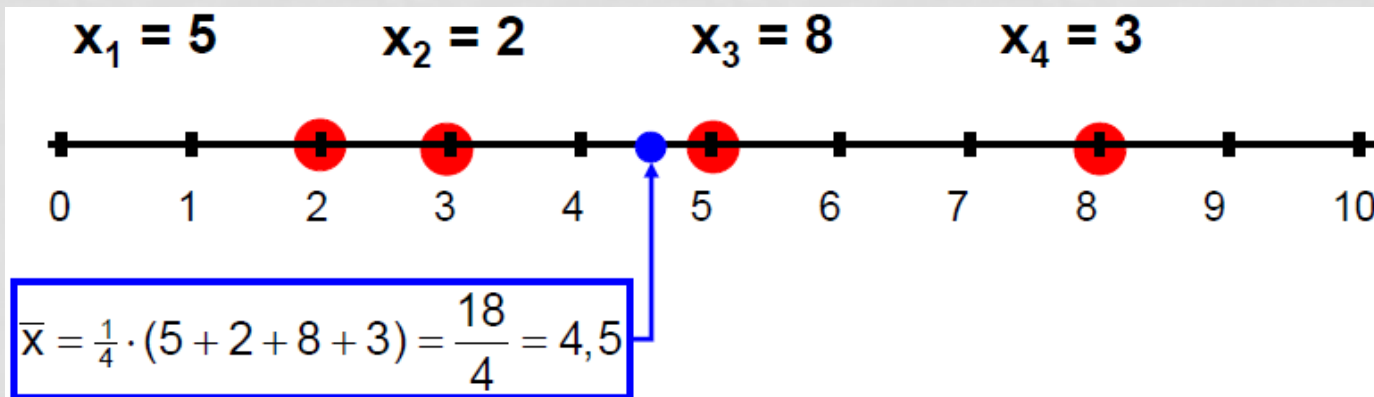
$$s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$



Wir untersuchen in unseren Beispielen nur den Bezug zur Grundgesamtheit und benutzen nur diese Formel!

VARIANZ

Beispiel:



Berechnung der Varianz

$$s^2 = \frac{1}{4} \cdot ((5 - 4,5)^2 + (2 - 4,5)^2 + (8 - 4,5)^2 + (3 - 4,5)^2) =$$
$$\frac{1}{4} \cdot (0,25 + 6,25 + 12,25 + 2,25) = \frac{21}{4} = 5,25$$

VARIANZ AUS HÄUFIGKEITSTABELLEN

Formel (1):

$$s^2 = \frac{1}{n} \sum_{i=1}^j (x_i - \bar{x})^2 * h(x_i) = \sum_{i=1}^j (x_i - \bar{x})^2 * f(x_i)$$

Formel (2):

$$s^2 = \left(\frac{1}{n} \sum_{i=1}^j x_i^2 * h(x_i) \right) - \bar{x}^2 = \left(\sum_{i=1}^j x_i^2 * f(x_i) \right) - \bar{x}^2$$

x_1, \dots, x_j

Merkmalsausprägungen

$h(x_1), \dots, h(x_j)$

absolute Häufigkeiten

$f(x_1), \dots, f(x_j)$

relative Häufigkeiten

j

Anzahl der Merkmalsausprägung x_i

VARIANZ AUS HÄUFIGKEITSTABELLEN

Berechnung der Varianz aus einer Häufigkeitstabelle nach **Formel (1)**:

Fall 1: Absolute Häufigkeit h_i

$$n = \sum_{i=1}^j h_i = h_1 + h_2 + \dots + h_j$$

$$s^2 = \frac{1}{n} \sum_{i=1}^j (x_i - \bar{x})^2 * h_i = \frac{1}{n} * ((x_1 - \bar{x})^2 h_1 + (x_2 - \bar{x})^2 h_2 + \dots + (x_j - \bar{x})^2 h_j)$$

h_i absolute Häufigkeit der Merkmalsausprägung x_i

n Summe der absoluten Häufigkeiten

j Anzahl der Merkmalsausprägung x_i

VARIANZ AUS HÄUFIGKEITSTABELLEN

Berechnung der Varianz aus einer Häufigkeitstabelle nach **Formel (1)**:

Fall 2: Relative Häufigkeit f_i

$$s^2 = \sum_{i=1}^j (x_i - \bar{x})^2 * f_i = ((x_1 - \bar{x})^2 f_1 + (x_2 - \bar{x})^2 f_2 + \dots + (x_j - \bar{x})^2 f_j)$$

f_i *relative Häufigkeit der Merkmalsausprägung x_i*

n *Summe der absoluten Häufigkeiten*

j *Anzahl der Merkmalsausprägung x_i*

VARIANZ AUS HÄUFIGKEITSTABELLEN

Beispiel:

Häufigkeitstabelle

Note x_i	1	2	3	4	5	6
Anzahl Schüler h_i	5	8	14	16	5	2
Relative Häufigkeit $f_i=h_i/n$	0,1	0,16	0,28	0,32	0,1	0,04

Schüler insgesamt:

$$n = \sum_{i=1}^6 h_i = 5 + 8 + 14 + 16 + 5 + 2 = 50$$

VARIANZ AUS HÄUFIGKEITSTABELLEN

Beispiel:

Berechnung der Varianz über die absolute Häufigkeit:

i	x_i	h_i	$x_i h_i$	\bar{x}	$x_i - \bar{x}$	$(x_i - \bar{x})^2 h_i$
1	1	5	5	3,28	-2,28	25,992
2	2	8	16	3,28	-1,28	13,107
3	3	14	42	3,28	-0,28	1,098
4	4	16	64	3,28	0,72	8,294
5	5	5	25	3,28	1,72	14,792
6	6	2	12	3,28	2,72	14,797
Σ		50	164	$\bar{x} = 164/50 = 3,28$		78,08

$$s^2 = \frac{1}{50} \sum_{i=1}^6 (x_i - \bar{x})^2 \cdot h_i = \frac{78,08}{50} = 1,562$$

VARIANZ AUS HÄUFIGKEITSTABELLEN

Beispiel:

Berechnung der Varianz über die relative Häufigkeit:

i	x_i	h_i	f_i	$x_i f_i$	\bar{x}	$x_i - \bar{x}$	$(x_i - \bar{x})^2 f_i$
1	1	5	0,1	0,1	3,28	-2,28	0,520
2	2	8	0,16	0,32	3,28	-1,28	0,262
3	3	14	0,28	0,84	3,28	-0,28	0,022
4	4	16	0,32	1,28	3,28	0,72	0,166
5	5	5	0,1	0,50	3,28	1,72	0,296
6	6	2	0,04	0,24	3,28	2,72	0,296
Σ		50	1	$\bar{x}=3,28$			$s^2=1,562$

$$s^2 = \sum_{i=1}^6 (x_i - \bar{x})^2 \cdot f_i = 1,562$$

VARIANZ AUS HÄUFIGKEITSTABELLEN

Berechnung der Varianz aus einer klassierten Häufigkeitstabelle nach **Formel (1)**:

Fall 1: Absolute Häufigkeit h_i

$$n = \sum_{i=1}^k h_i = h_1 + h_2 + \dots + h_k$$

$$s^2 = \frac{1}{n} \sum_{i=1}^k \underbrace{(m_i - \bar{x})^2}_{\text{}} * h_i = \frac{1}{n} * ((m_1 - \bar{x})^2 h_1 + (m_2 - \bar{x})^2 h_2 + \dots + (m_k - \bar{x})^2 h_k)$$

h_i absolute Häufigkeit der i -ten Klasse

n Summe der absoluten Häufigkeiten

k Anzahl der Klassen

m_i Klassenmitte der i -ten Klasse

VARIANZ AUS HÄUFIGKEITSTABELLEN

Berechnung der Varianz aus einer klassierten Häufigkeitstabelle nach **Formel (1)**:

Fall 2: Relative Häufigkeit f_i

$$s^2 = \sum_{i=1}^k (m_i - \bar{x})^2 * f_i = ((m_1 - \bar{x})^2 f_1 + (m_2 - \bar{x})^2 f_2 + \dots + (m_k - \bar{x})^2 f_k)$$

f_i	<i>relative Häufigkeit der i-ten Klasse</i>
n	<i>Summe der absoluten Häufigkeiten</i>
k	<i>Anzahl der Klassen</i>
m_i	<i>Klassenmitte der i-ten Klasse</i>

VARIANZ AUS HÄUFIGKEITSTABELLEN

Beispiel:

klassierte Häufigkeitstabelle für die Körpergröße:

Klasse x_i	150 b. u. 160	160 b. u. 170	170 b. u. 180	180 b. u. 190
Häufigkeit h_i	9	12	7	2
Klassenmitte m_i	155	165	175	185
Relative Häufigkeit $f_i=h_i/n$	0,3	0,4	0,23	0,07

Schüler insgesamt:

$$n = \sum_{i=1}^4 h_i = 9 + 12 + 7 + 2 = 30$$

VARIANZ AUS HÄUFIGKEITSTABELLEN

Beispiel:

Berechnung der Varianz über die absolute Häufigkeit:

i	Klasse x_i	m_i	h_i	$m_i h_i$	\bar{x}	$m_i - \bar{x}$	$(m_i - \bar{x})^2 h_i$
1	150 b. u. 160	155	9	1.392	165,67	-10,67	1.024,64
2	160 b. u. 170	165	12	1.980	165,67	-0,67	5,39
3	170 b. u. 180	175	7	1.225	165,67	9,33	609,34
4	180 b. u. 190	185	2	370	165,67	19,33	747,30
Σ			30	4.970	$\bar{x}=4.970/30=165,67$		2.386,67

$$s^2 = \frac{1}{30} \sum_{i=1}^6 (m_i - \bar{x})^2 * h_i = \frac{2.386,67}{30} \approx 80$$

VARIANZ AUS HÄUFIGKEITSTABELLEN

Beispiel:

Berechnung der Varianz über die relative Häufigkeit:

i	Klasse x_i	m_i	h_i	f_i	$m_i f_i$	\bar{x}	$m_i - \bar{x}$	$(m_i - \bar{x})^2 f_i$
1	150 b. u. 160	155	9	0,3	46,5	165,67	-10,67	34,1547
2	160 b. u. 170	165	12	0,4	66,0	165,67	-0,67	0,1796
3	170 b. u. 180	175	7	0,23	40,25	165,67	9,33	20,0212
4	180 b. u. 190	185	2	0,07	12,95	165,67	19,33	26,1554
Σ			30	1	$\bar{x}=165,67$			80,51

$$s^2 = \sum_{i=1}^6 (m_i - \bar{x})^2 * f_i \approx 80$$

STANDARDABWEICHUNG

Standardabweichung:

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

Die Standardabweichung ist ein Maß dafür, wie hoch die Aussagekraft des Mittelwertes ist. Eine kleine Standardabweichung bedeutet, alle Beobachtungswerte liegen nahe am Mittelwert (kleine Streuung).

Eine große Standardabweichung bedeutet, die Beobachtungswerte sind weit um den Mittelwert gestreut.

bei normalverteilten Daten liegen ca. 95% der Beobachtungswerte im Intervall $[\bar{x} - 2s, \bar{x} + 2s]$.

STREUUNGSPARAMETER

Spannweite w :

$$w = x_{max} - x_{min}$$

(Inter)Quartilsabstand:

$$Q_A = IQR = Q_3 - Q_1$$

Varianz:

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

Standardabweichung:

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

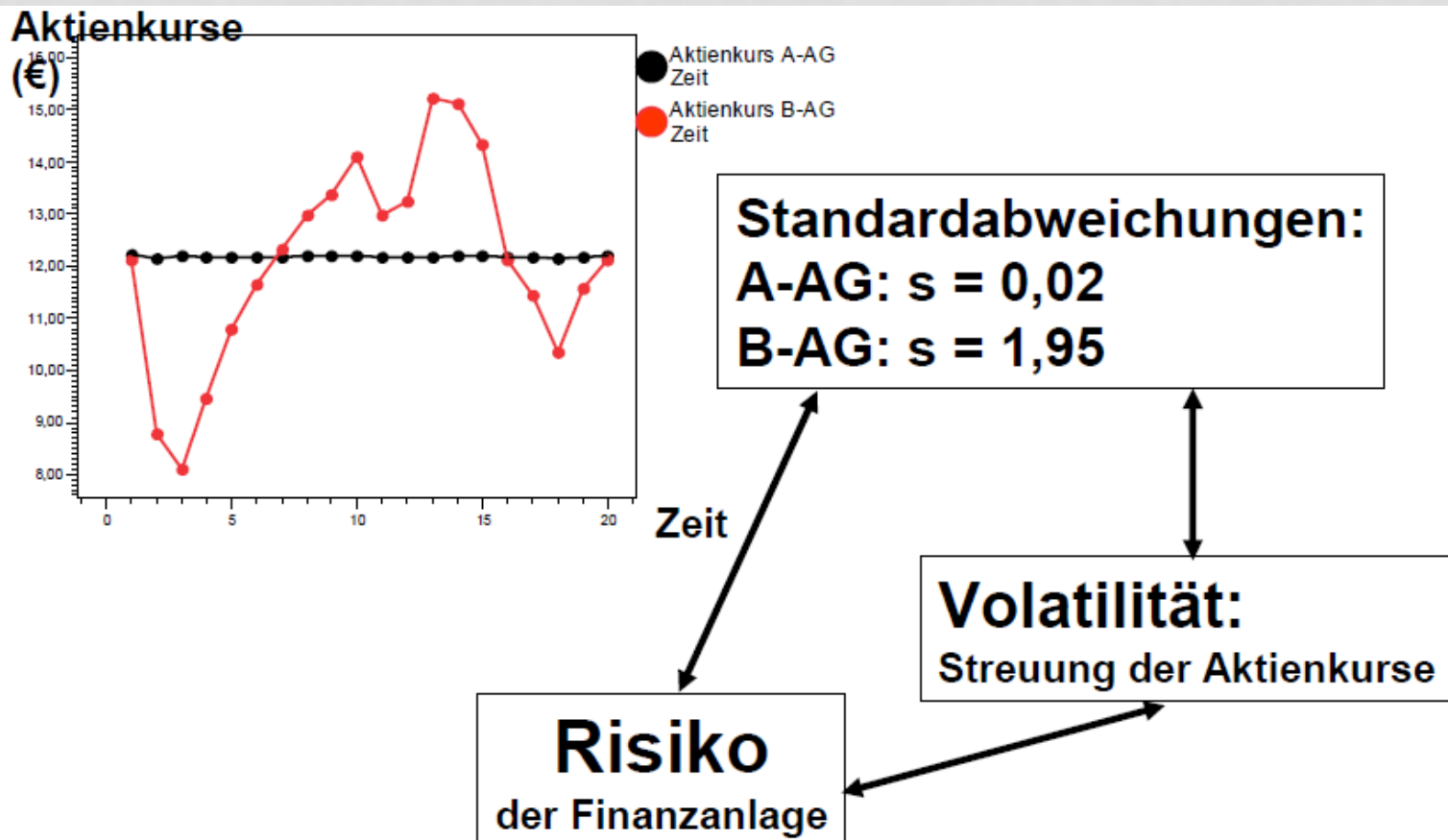
Variationskoeffizient:

$$v = \frac{s}{\bar{x}}$$

(dimensionslose Größe)

STREUUNGSPARAMETER

Beispiel für die Anwendung:



„EINIGE STATISTIKBEGRIFFE IN ENGLISCH“

deutsch

Grundgesamtheit

Stichprobe

arithmetisches Mittel

Modus

Spannweite

Varianz

Standardabweichung

englisch

population

sample

mean

mode

range

variance

standard deviation

(std dev)