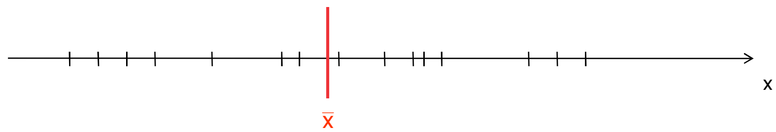
Stichprobe

x_1
x_2
x_3
$\vdots$
x_n
n

Lageparameter

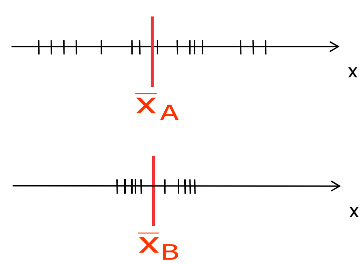
Arithmetischer Mittelwert

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



Stichprobe A

x_1
x_2
$\vdots$
x_{n_A}
n_A



Stichprobe B

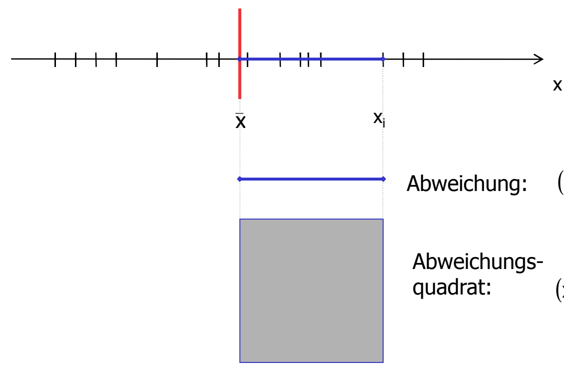
x_1
x_2
$\vdots$
x_{n_B}
n_B

→ Mittelwert allein reicht nicht zur Beschreibung einer Stichprobe

Suche nach Maß für die Streuung

Stichprobe

x_1
x_2
$\vdots$
x_n
n
$\bar{x}$



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

Abweichungs-quadrat: $(x_i - \bar{x})^2$

universitätbonn
Deskriptive Statistik: Statistische Maßzahlen
IMBIE

Streuungsparameter

Stichprobe

x_1
x_2
$\vdots$
x_n
n
$\bar{x}$

Abweichungs-quadrate

$(x_1 - \bar{x})^2$
$(x_2 - \bar{x})^2$
$\vdots$
$(x_n - \bar{x})^2$

Mittelwert der Abweichungsquadrate

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

Empirische Varianz

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

Deskriptive Statistik: Statistische Maßzahlen

universitätbonn

IMBIE

Was geht in die Formel für s^2 ein?

x_1
x_2
$\vdots$
x_n
n
$\bar{x}$

Beispiel:

$$\begin{aligned} x_1 &= 75 \\ x_2 &= 2 \\ x_3 &= 270 \\ x_4 &= 4 \cdot 100 - 75 - 2 - 270 = 53 \\ n &= 4 \\ \bar{x} &= 100 \end{aligned}$$

x_4 ist nicht mehr frei wählbar,
sondern wird durch die anderen Werte festgelegt

→ s^2 hat (n-1) Freiheitsgrade (f)

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

→ $s = \sqrt[+]{s^2}$ **Empirische Standardabweichung**

Deskriptive Statistik: Statistische Maßzahlen

universitätbonn

IMBIE

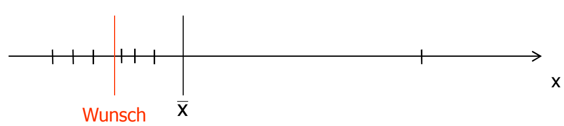
Lageparameter	Streuungsparameter
arithmetischer Mittelwert $\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$	***empirische Varianz*** $s^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2$ ***empirische Standardabweichung*** $s = \sqrt[+]{s^2}$

Deskriptive Statistik: Statistische Maßzahlen
Ausreißer

<u>Frage: Wie werden Ausreißer behandelt?</u>

1) Wegschmeißen	NEIN!
2) Wert überprüfen und wenn begründet korrigieren	Ja!
3) Zweimal Mittelwert berechnen und Vorgehen beschreiben	
4) Inhomogene Stichprobe	
5) Ausreißer wird bestätigt	Was nun???

Deskriptive Statistik: Statistische Maßzahlen
Ausreißer



„Wunsch" basiert auf der Reihenfolge der Einzelwerte

„Wunsch"-Position liegt in der „Mitte" der Werte

Dazu muss man die Einzelwerte der Größe nach sortieren.

Deskriptive Statistik: Statistische Maßzahlen
Rangliste

IMBIE

Urliste	Rangliste
$x_1=5$	$x_{(1)}=3$
$x_2=9$	$x_{(2)}=4$
$x_3=3$	$x_{(3)}=5$
$x_4=8$	$x_{(4)}=6$
$x_5=19$	$x_{(5)}=8$
$x_6=4$	$x_{(6)}=9$
$x_7=6$	$x_{(7)}=19$
$n=7$	$n=7$

Parameter aus der Rangliste:

Minimum:	$\min = x_{(1)}$
Maximum:	$\max = x_{(n)}$
Spannweite:	$R = x_{(n)} - x_{(1)}$

Deskriptive Statistik: Statistische Maßzahlen

IMBIE

Urliste	Rangliste
$x_1=5$	$x_{(1)}=3$
$x_2=9$	$x_{(2)}=4$
$x_3=3$	$x_{(3)}=5$
$x_4=8$	$x_{(4)}=6$
$x_5=19$	$x_{(5)}=8$
$x_6=4$	$x_{(6)}=9$
$x_7=6$	$x_{(7)}=19$
$n=7$	$n=7$

Lageparameter → in die Mitte

Lageparameter soll höchstens die Hälfte der Rangliste überschreiten und höchstens die Hälfte der Rangliste unterschreiten

Welcher Wert erfüllt diese Bedingung? $x_{(4)} =: \tilde{x}$ Median

Wie kommt man bei n=7 Rangwerten zu $x_{(4)}$ als Median?
Indem man rechnet: (7+1)/2=4

Allgemein: Bei n Rangwerten ist der Median definiert als $\boxed{\tilde{x} := x_{\left(\frac{n+1}{2}\right)}}$

Funktioniert nur bei ungeraden n !

universitätbonn

Deskriptive Statistik: Statistische Maßzahlen

IMBIE

Was macht man bei n gerade?

Urliste	Rangliste
$x_1=5$	$x_{(1)}=3$
$x_2=9$	$x_{(2)}=4$
$x_3=3$	$x_{(3)}=5$
$x_4=8$	$x_{(4)}=6$
$x_5=19$	$x_{(5)}=7$
$x_6=4$	$x_{(6)}=8$
$x_7=6$	$x_{(7)}=9$
$X_8=7$	$X_{(8)}=19$

Die Vorgabe für den Median wird in diesem Beispiel erfüllt von $x_{(4)}$ aber auch von $x_{(5)}$.

Was macht man?

Bilde den Mittelwert aus beiden Werten $\tilde{x} := \frac{x_{(4)} + x_{(5)}}{2}$

Allgemein bei n gerade $\boxed{\tilde{x} := \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2}}$

Der Median ist unempfindlich gegen Ausreißer.

universitätbonn

Deskriptive Statistik: Statistische Maßzahlen

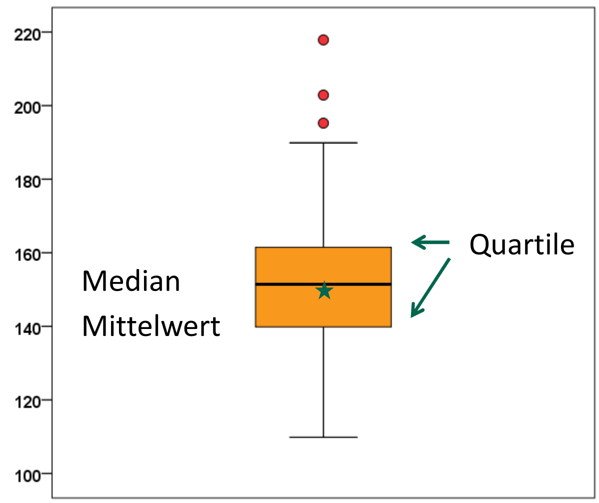
IMBIE

Lageparameter	Streuungsparameter
Mittelwert $\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$	***empirische Varianz*** $s^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2$ ***emp. Standardabweichung*** $s = \sqrt[+]{s^2}$
Median $\tilde{x} = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{für n ungerade} \\ \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2} & \text{für n gerade} \end{cases}$	***Spannweite*** $R = x_{(n)} - x_{(1)}$

universitätbonn

Deskriptive Statistik: Boxplot

IMBIE



universitätbonn

Bivariate Statistik: Kontingenztafel

IMBIE

Betrachte 2 qualitative Merkmale A und B je Beobachtungseinheit

Merkmal A hat Ausprägungen $a_1, a_2, \ldots, a_k$
Merkmal B hat Ausprägungen $b_1, b_2, \ldots, b_l$

		Merkmal B						
		b_1	b_2	…	b_j	…	b_l	Σ
Merkmal A	a_1	n_{11}	n_{12}	…	n_{1j}	…	n_{1l}	$n_{1.}$
	a_2	n_{21}	n_{22}	…	n_{2j}	…	n_{2l}	$n_{2.}$
	⋮	⋮	⋮	…	⋮	…	⋮	⋮
	a_i	n_{i1}	n_{i2}	…	n_{ij}	…	n_{il}	$n_{i.}$
	⋮	⋮	⋮	…	⋮	…	⋮	⋮
	a_k	n_{k1}	n_{k2}	…	n_{kj}	…	n_{kl}	$n_{k.}$
	Σ	$n_{.1}$	$n_{.2}$	…	$n_{.j}$	…	$n_{.l}$	n

Kontingenztafel
(k × l – Felder- Tafel)

absolute Häufigkeit n_{ij}: Anzahl der Beobachtungen, die gleichzeitig die Ausprägungen a_i und b_j annehmen

Randhäufigkeiten: $n_{i.} = n_{i1} + n_{i2} + \ldots + n_{il}$ $\quad i = 1, \ldots, k$
$n_{.j} = n_{1j} + n_{2j} + \ldots + n_{kj}$ $\quad j = 1, \ldots, l$

Beispiel: *Klausurergebnisse nach Geschlecht*

			Ergebnis: durchgefallen	Ergebnis: bestanden	Gesamt
Geschlecht	Männer	Anzahl	20	19	39
		% von Geschlecht	51.3%	48.7%	100.0%
		% von Ergebnis	57.1%	55.9%	56.5%
		% der Gesamtzahl	29.0%	27.5%	56.5%
	Frauen	Anzahl	15	15	30
		% von Geschlecht	50.0%	50.0%	100.0%
		% von Ergebnis	42.9%	44.1%	43.5%
		% der Gesamtzahl	21.7%	21.7%	43.5%
Gesamt		Anzahl	35	34	69
		% von Geschlecht	50.7%	49.3%	100.0%
		% von Ergebnis	100.0%	100.0%	100.0%
		% der Gesamtzahl	50.7%	49.3%	100.0%

$$\sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij} = \sum_{i=1}^{k} n_{i.} = \sum_{j=1}^{l} n_{.j} = n$$

Relative Häufigkeiten: $h_{ij} = \frac{n_{ij}}{n}$ mit i=1,...,k und j=1,...,l $\left(\sum_{i=1}^{k}\sum_{j=1}^{l} h_{ij} = 1\right)$

Spaltenprozente: $\frac{n_{ij}}{n_{.j}}$ $\left(\sum_{i=1}^{k} \frac{n_{ij}}{n_{.j}} = 1\right)$

Zeilenprozente: $\frac{n_{ij}}{n_{i.}}$ $\left(\sum_{j=1}^{l} \frac{n_{ij}}{n_{i.}} = 1\right)$

mit i=1,...,k und j=1,...,l