

# Deskriptiver Statistik

Datenzusammenfassung

G.Schay

# Variabilität



Es gibt eine eingebaute unsicherheit und Variabilität in der Natur

das erfordert ein etwas anderes Denken

# Statistik beschreibt hauptsächlich zufällige Massenerscheinungen



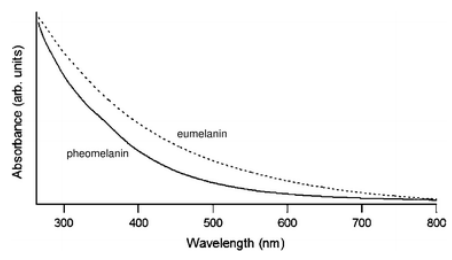
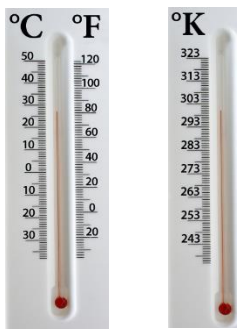
- *Datensammlung (Stichprobe)*
  - *Datenorganisation*
  - *Datenanalyse*
- *Konklusion, Entscheidung*

*Deskriptiver Statistik*

*Induktiver Statistik*



# Merkmal, Beobachtung





# Merkmalstype












qualitativ  
kategorisch

= ≠


→

= ≠  
< >




			
Blue	Turquoise	Amethyst	Gray
			
Green	Honey	Pure Hazel	True Sapphire
			
*New Brilliant Blue	*New Sterling Gray	*New Gemstone Green	


1 TALC




2 GYPSUM




3 CALCITE




4 FLOURITE




5 APATITE




6 FELDSPAR




7 QUARTZ




8 TOPAZ



9 CORUNDUM



10 DIAMOND




quantitativ  
numerisch


intervallskala

verhältnisskala


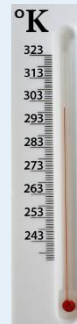
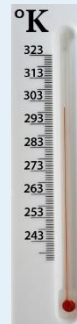
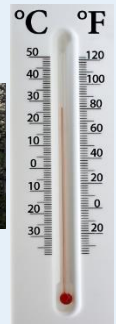

= ≠  
< >  
+ -



= ≠  
< >  
+ -



absoluter  
Nullpunkt



# Nominal: wir haben Namen

Beobachtungsliste

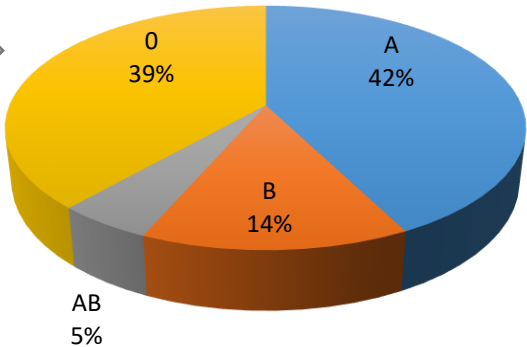
patient №	blood group (ABO)	cholesterol level (mg/dL)
1	B	148
2	AB	147
3	B	169
4	B	159
5	B	150
6	B	167
7	A	144
8	B	158
9	AB	177

Häufigkeitsverteilung

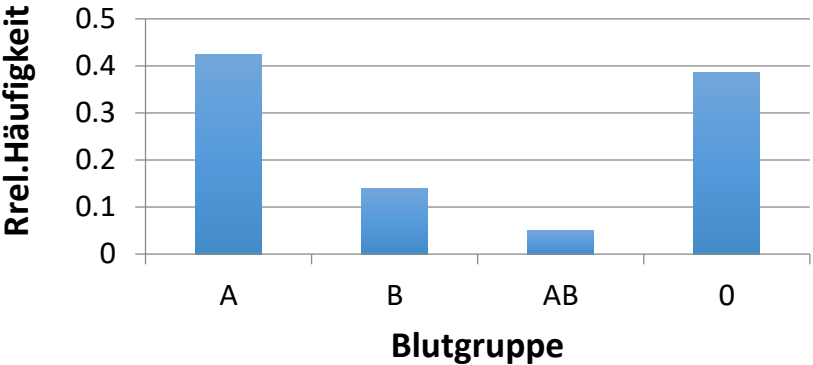
blood group	(absolute) frequency	relative frequency
A	85	0.425
B	28	0.14
AB	10	0.05
O	77	0.385
Σ	200	1

grafische Darstellung

Relative Häufigkeiten



Säulendiagramm



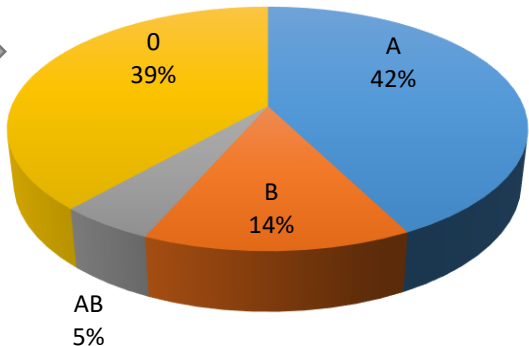
# Nominal: wir haben Namen

Beobachtungsliste

patient №	blood group (ABO)	cholesterol level (mg/dL)
1	B	148
2	AB	147
3	B	169
4	B	159
5	B	150
6	B	167
7	A	144
8	B	158
9	AB	177

grafische Darstellung

Relative Häufigkeiten



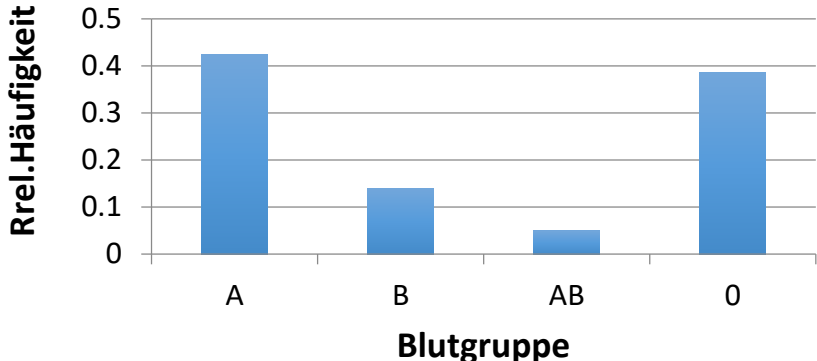
„Mitte?“

es gibt keine feste Anordnung oder Reihenfolge

häufigster Wert = Modalwert (Modus)

hier: „A“

Säulendiagramm



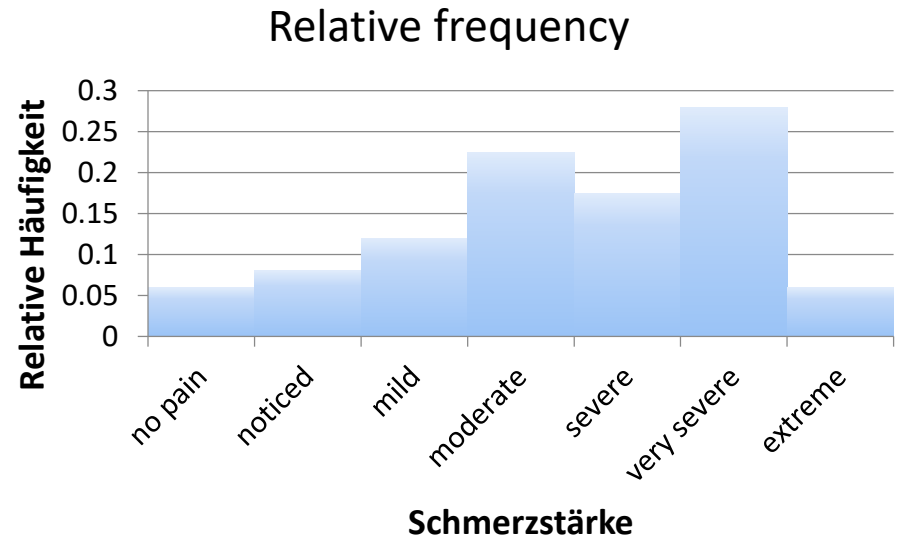
# ordinale Merkmale

Frequency table

Severity of pain	Relative frequency	Cumulative relative frequency
no pain	0,06	0,06
noticed	0,08	0,14
mild	0,12	0,26
moderate	0,225	0,485
severe	0,175	0,66
very severe	0,28	0,94
extreme	0,06	1
$\Sigma$	1	

„Mitte“

Modalwert ist hier (und immer) benutzbar.



Hier gibt es aber eine Reihenfolge -> die kumulierte Verteilung ist auch konstruierbar!  
Häufigkeit von  $\xi < X$ , also wie viele Beobachtungen ( $\xi$ ) sind KLEINER ALS ein Schwellenwert ( $X$ )

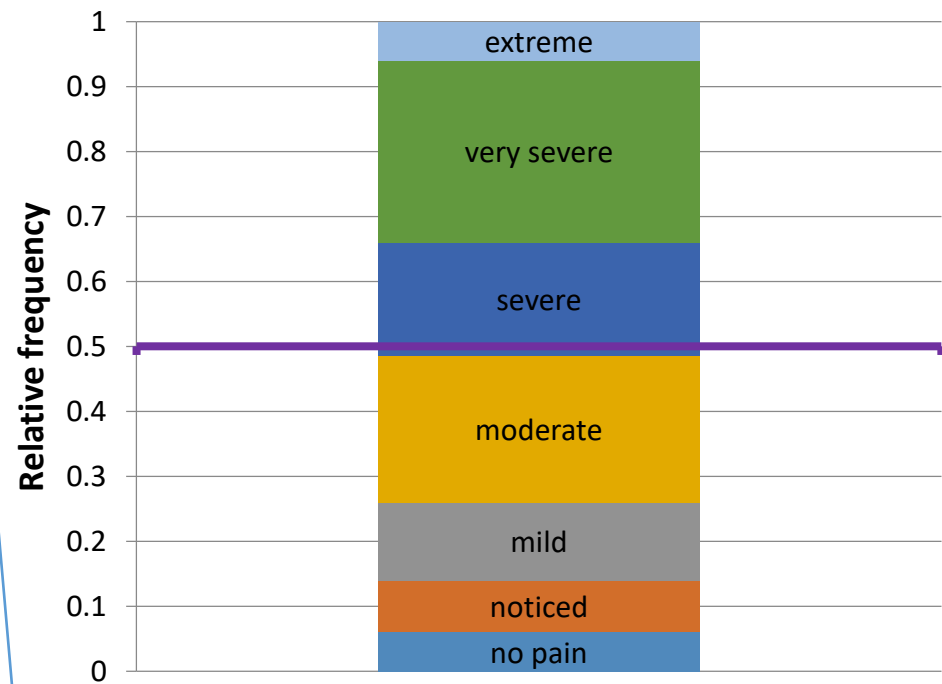
**MEDIAN** : der Wert, wovon 50% der Beobachtungen kleiner sind.



# Description of Ordinal Variables II.

Severity of pain (X)	Cumulative relative frequency
no pain	0,06
noticed	0,14
mild	0,26
moderate	0,485
severe	0,66
very severe	0,94
extreme	1
$\Sigma$	

Grafisch



wir suchen den X wozu 0.5 als relative kumulierte Häufigkeit gehört

# quantitative Merkmale

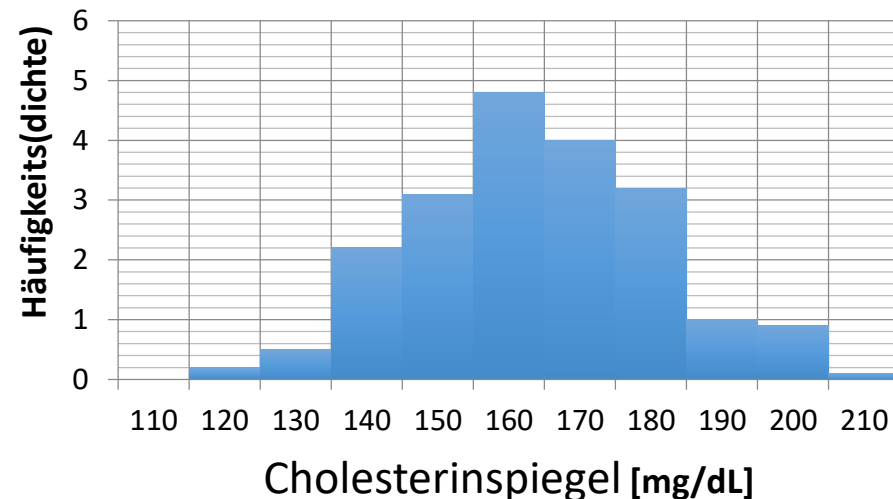
## Häufigkeitsverteilungen

## Grafisch

frequency distributions (differential discrimination functions)				
Klassen bins (classes, intervals)	abs.H. (absolute) frequency (FREQUENCY) <small>Häufigkeit</small>	rel.H. relative frequency	abs.H.dichte (absolute) frequency density	rel.H.dichte relative frequency density
$x \leq 100$	0			
$100 < x \leq 110$	0	0	0	0
$110 < x \leq 120$	2	0,01	0,2	0,001
$120 < x \leq 130$	5	0,025	0,5	0,0025
$130 < x \leq 140$	22	0,11	2,2	0,011
$140 < x \leq 150$	31	0,155	3,1	0,0155
$150 < x \leq 160$	48	0,24	4,8	0,024



## absolute Häufigkeitsverteilung



es gibt informationsverlust mit einem Säulendiagramm

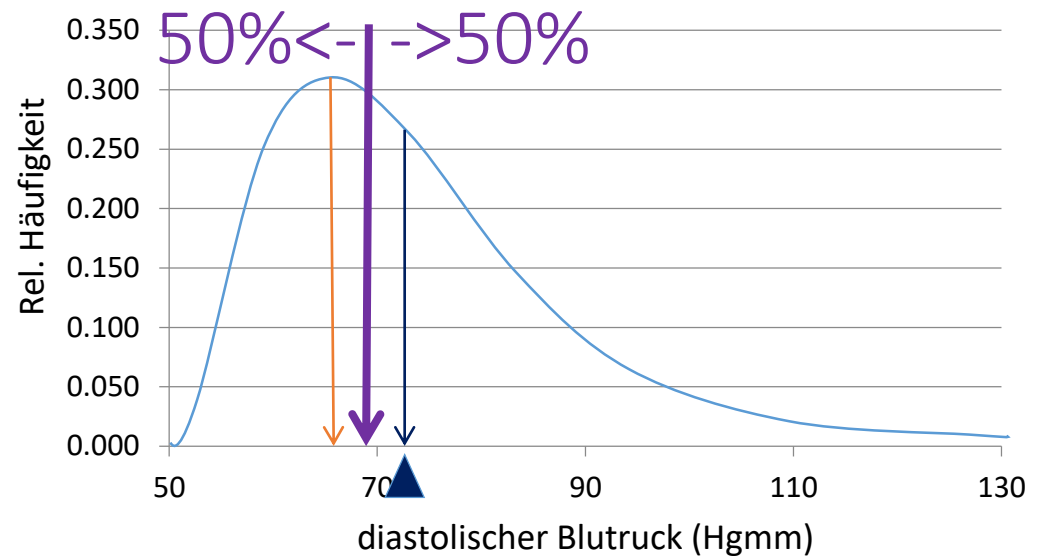
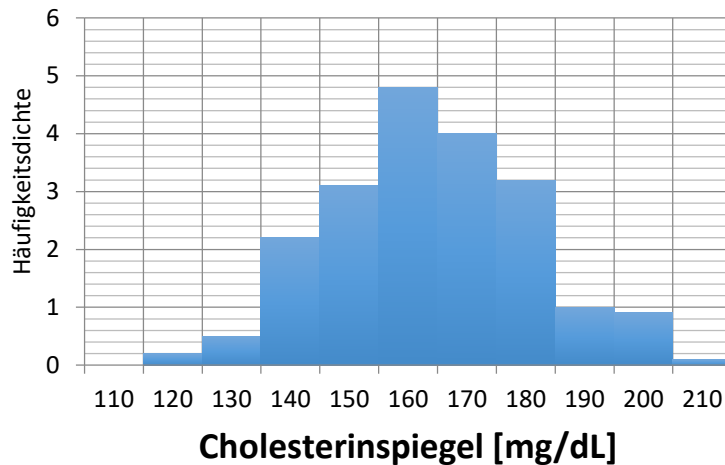
Klassenbreite:

praktische Hinsichte (wie sieht gut aus)

Statistisch?

-> **Kumuliert is eindeutig!**

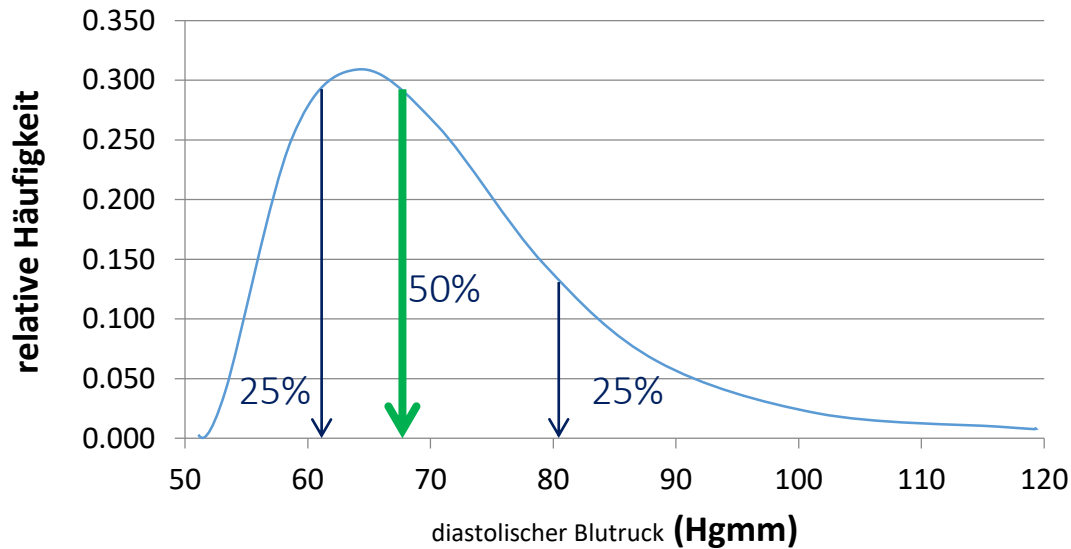
# quantitative Merkmale



## Lageparameter

- **Modalwert**: am häufigsten vorkommende
  - **Median**: „Mitte“
- **Mittelwert** (durchschnitt): „schwerepunkt“  $\bar{x}$

# Qantile (Perzentile)



- **Median:** 50% ( $Q_2$ )
- **Quartile:** untere Quartile ( $Q_1$ ): 25%; obere Quartile ( $Q_3$ ): 75%  
Generell: beliebige %-e sind möglich.

**Perzentile:** gegebener % ist zu links.

(ganz genau: max  $N \cdot p$  Beobachtungen sind kleiner, und maximal  $N \cdot (1-p)$  sind größer)

Day	Waiting time (min)		
1	1,27	median	8,48
2	3,3	lower quartile	3,59
3	3,44	mean	7,72
4	3,64		
5	6,33		
6	7,72		
7	9,23		
8	9,87		
9	10,31		
10	12,29		
11	12,3		
12	12,98		

Day	Waiting time (min)		
1	1,27	median	8,48
2	3,3	lower quartile	3,59
3	3,44	mean	8,31
4	3,64		
5	6,33		
6	7,72		
7	9,23		
8	9,87		
9	10,31		
10	12,29		
11	12,3		
12	20		

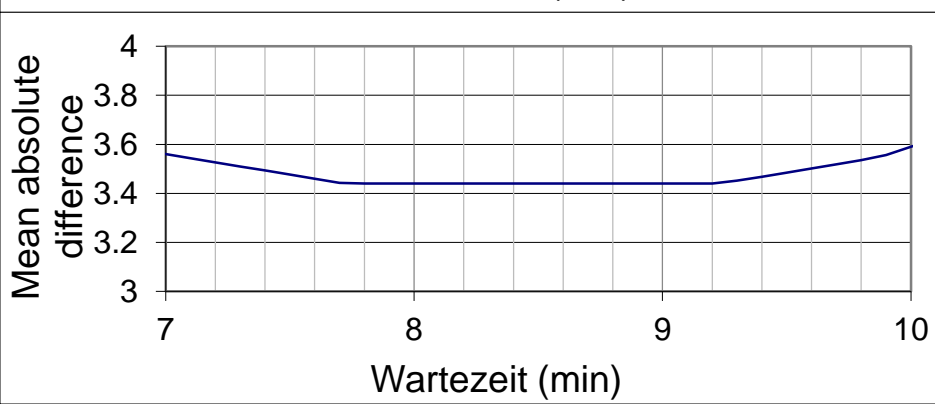
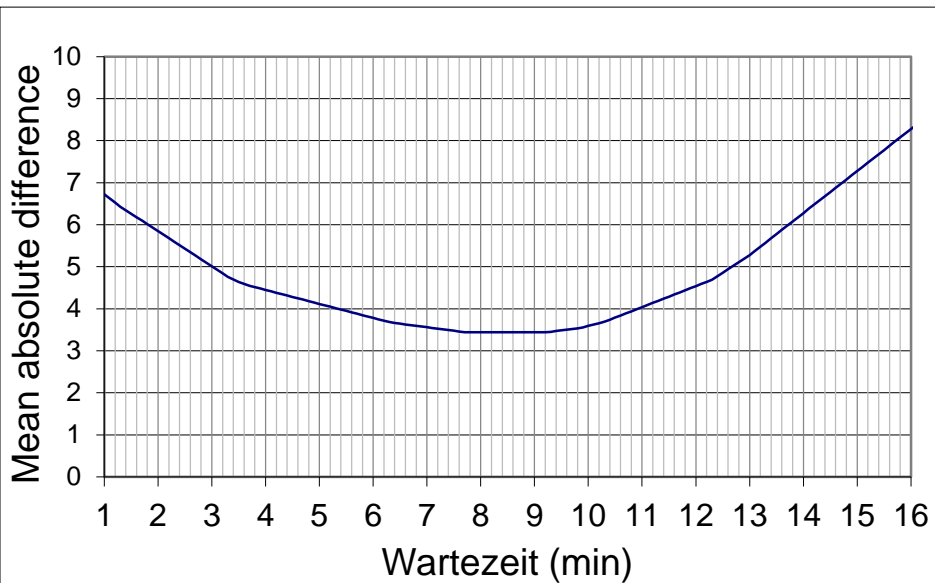
Ausreißer

Median und Quantile sind unempfindlich für Ausreißer, dagegen Mittelwert ist schon empfindlich!



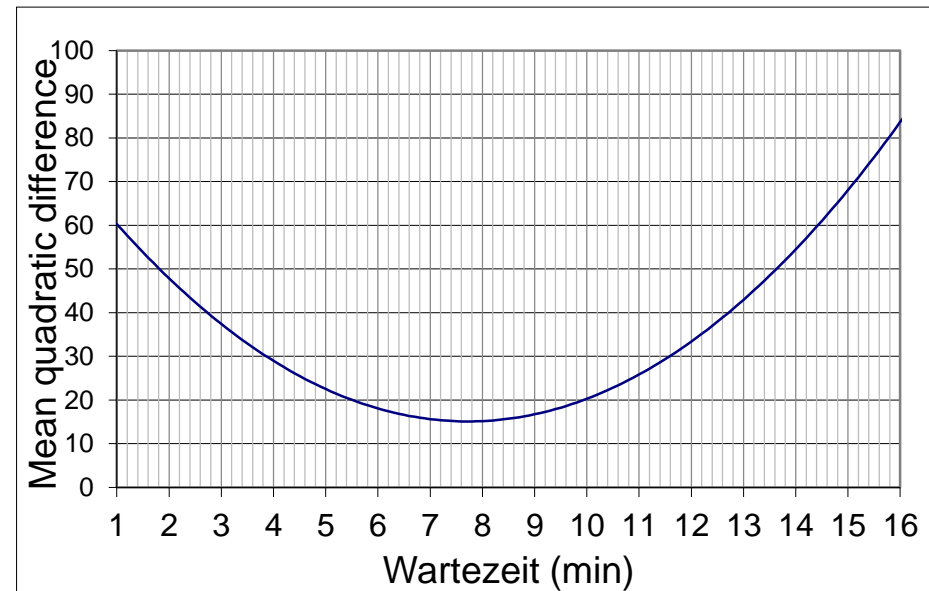
$\frac{1}{n} \sum |x_i - x^*|$  ist minimal wenn:

$x^* = \textit{Median}$

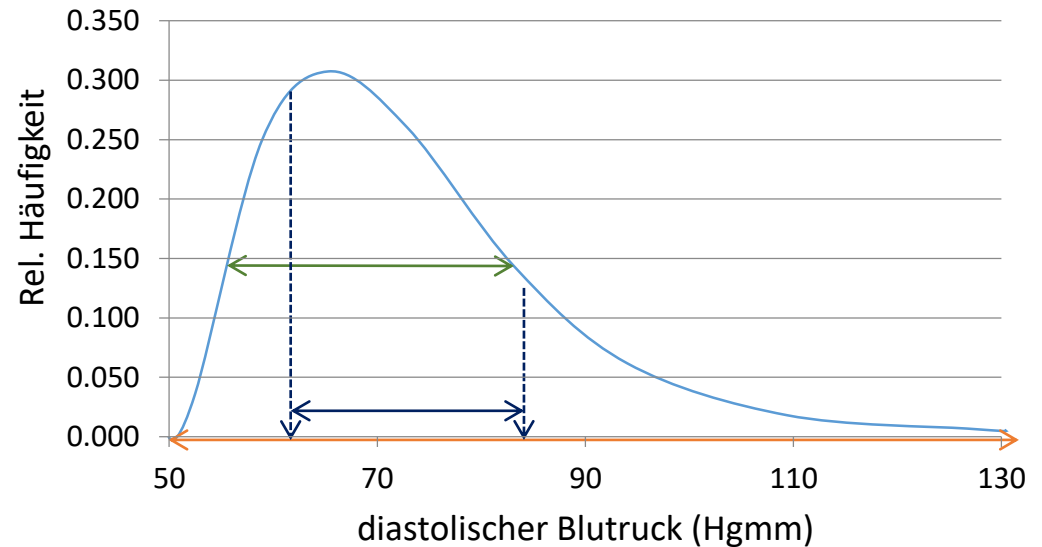
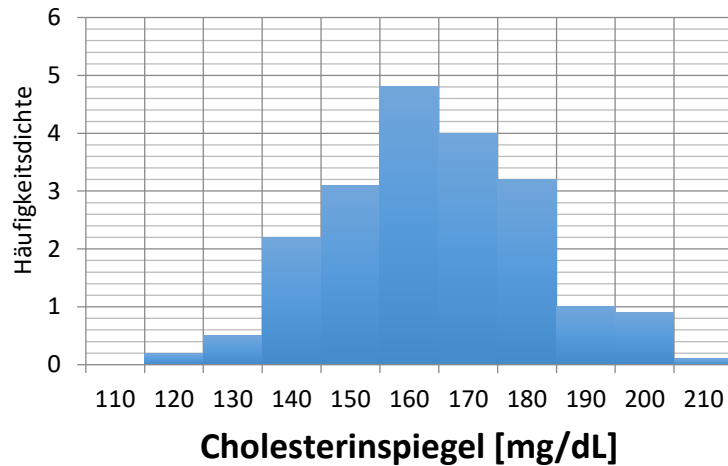


$\frac{1}{n} \sum (x_i - x^*)^2$  ist minimal wenn:

$x^* = \textit{Mittelwert}$



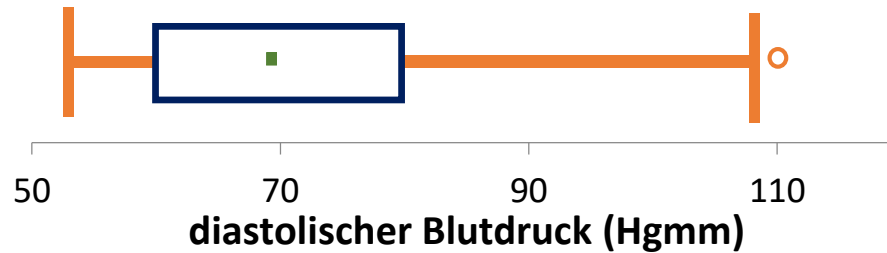
# Breiteparameter



## Breite-, Streuungs-parameter

- **volles Bereich**: max-min
- **Varianz ( $s^2$ )**: mittelwert der quadratischen Abweichungen vom Mittelwert (Korrigiert - Stichprobe, Unkorrigiert - Population)
  - **Standardabweichung ( $s, sd, SD$ )**: Wurzel der varianz
    - **Interquartiler Abstand (IQR)**:  $Q_{75\%} - Q_{25\%}$

# Box plot



Mittelpunkt: Mittelwert oder *Median*

Kästchen:  $2 \times$  Standardabweichung oder IQR

Schnurrhaare (*whisker*):  $3 \times$  SD ; minimum und maximum ; 0.05 and 0.95 Quantile,  $1.5 \times$  IQR...  
draussen: **Ausreißer**

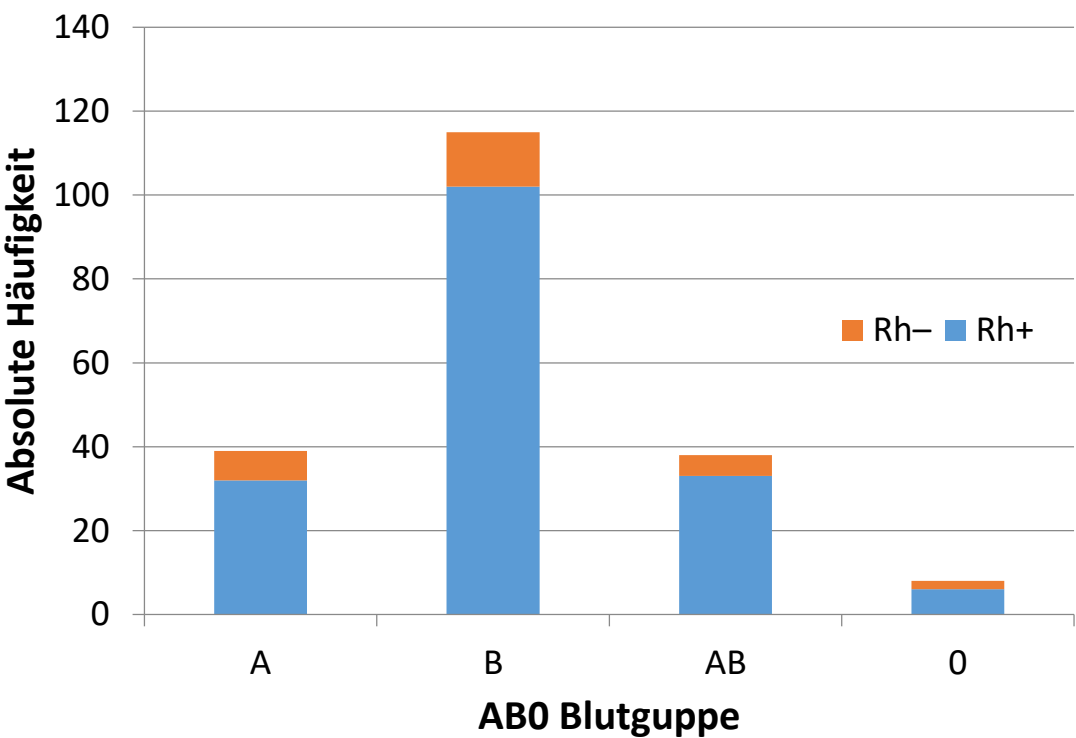
abgeschnittener Mittelwert: Mittelwert neugerechnet ohne Ausreißer.

# Qualitative Beschreibung mehrerer gemeinsamen Beobachtungen

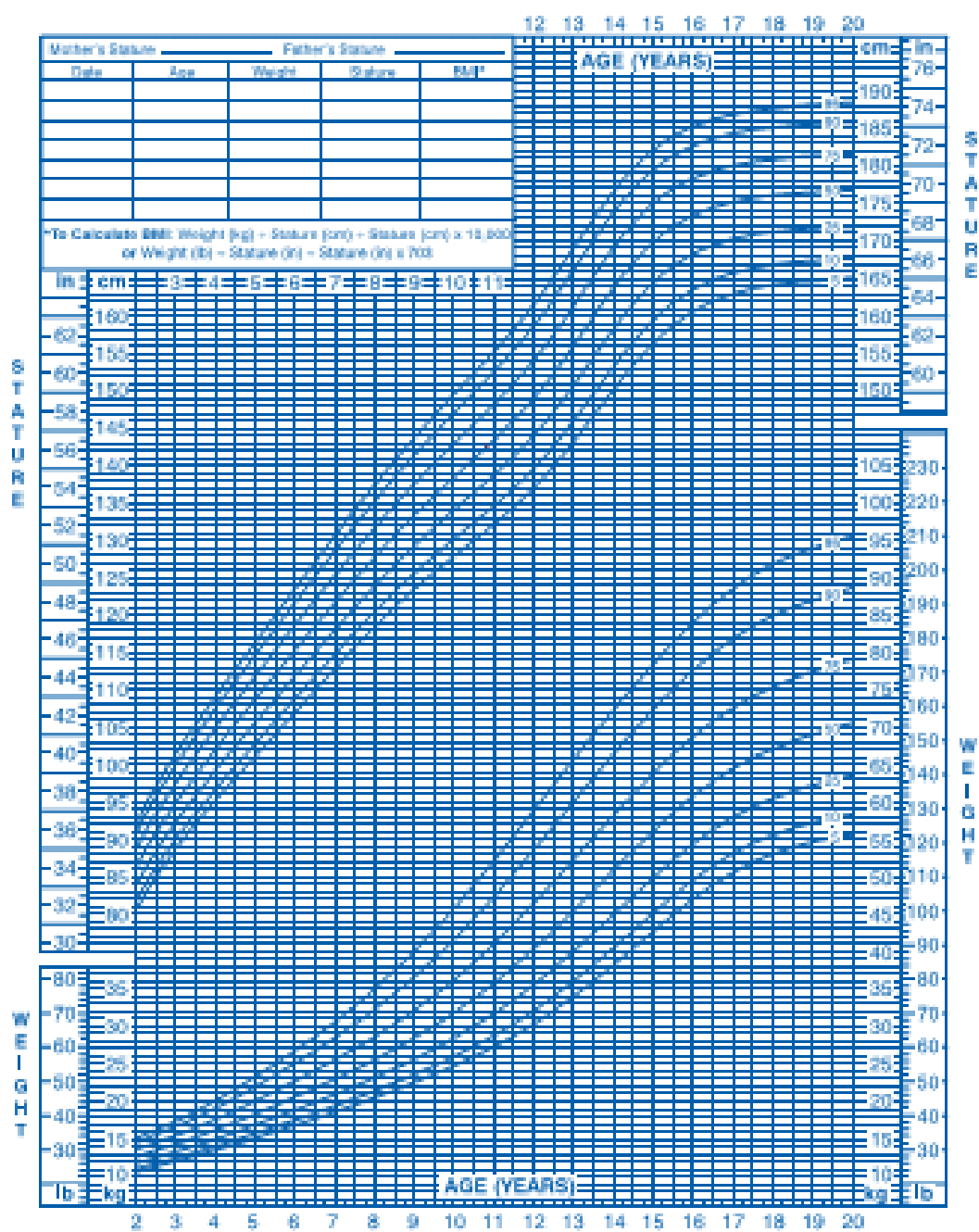
Kontingenztafel

	A	B	AB	0	$\Sigma$
Rh+	32	102	33	6	173
Rh-	7	13	5	2	27
$\Sigma$	39	115	38	8	200

gestapelter Säulendiagramm



# Perzentilkurven





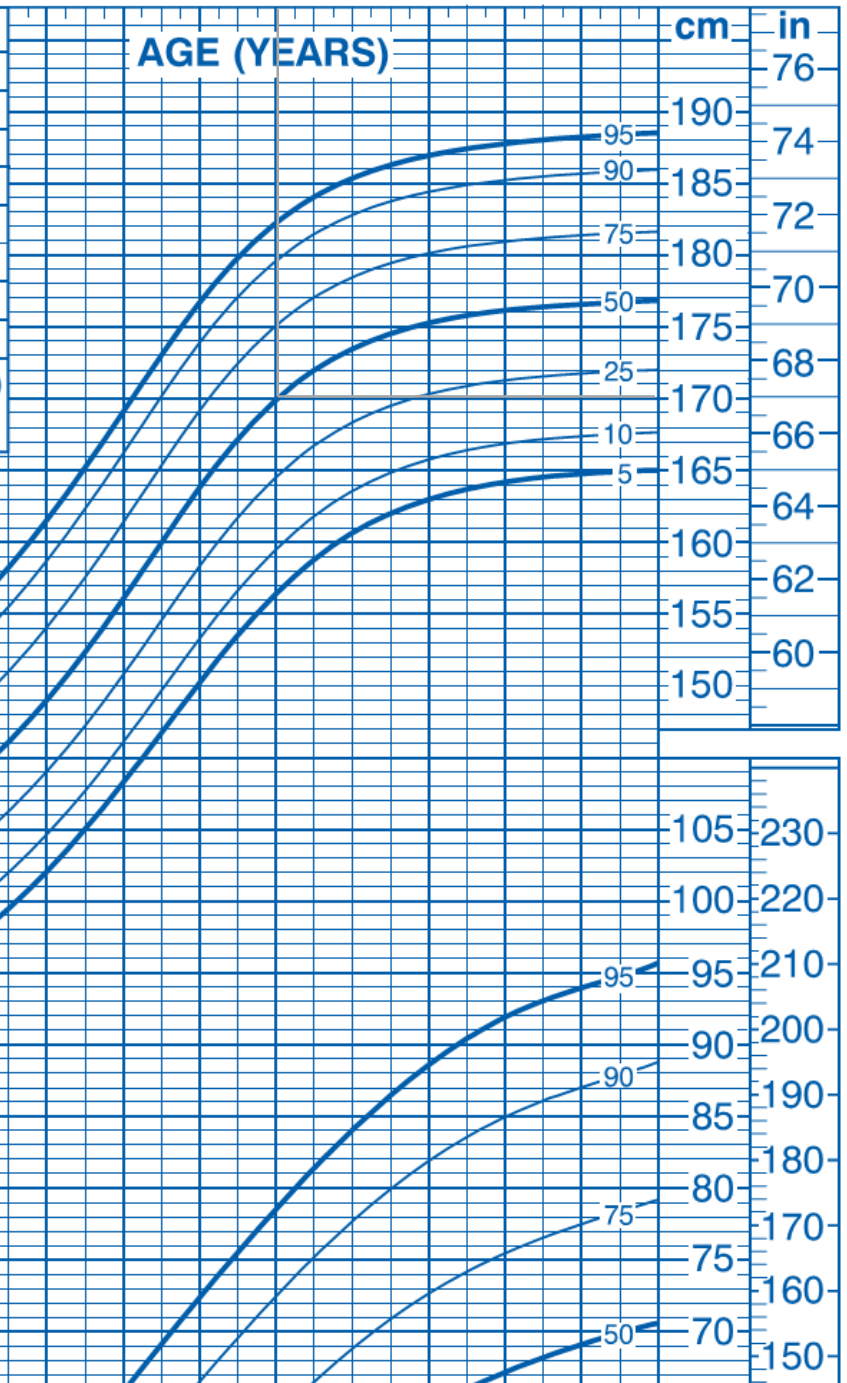
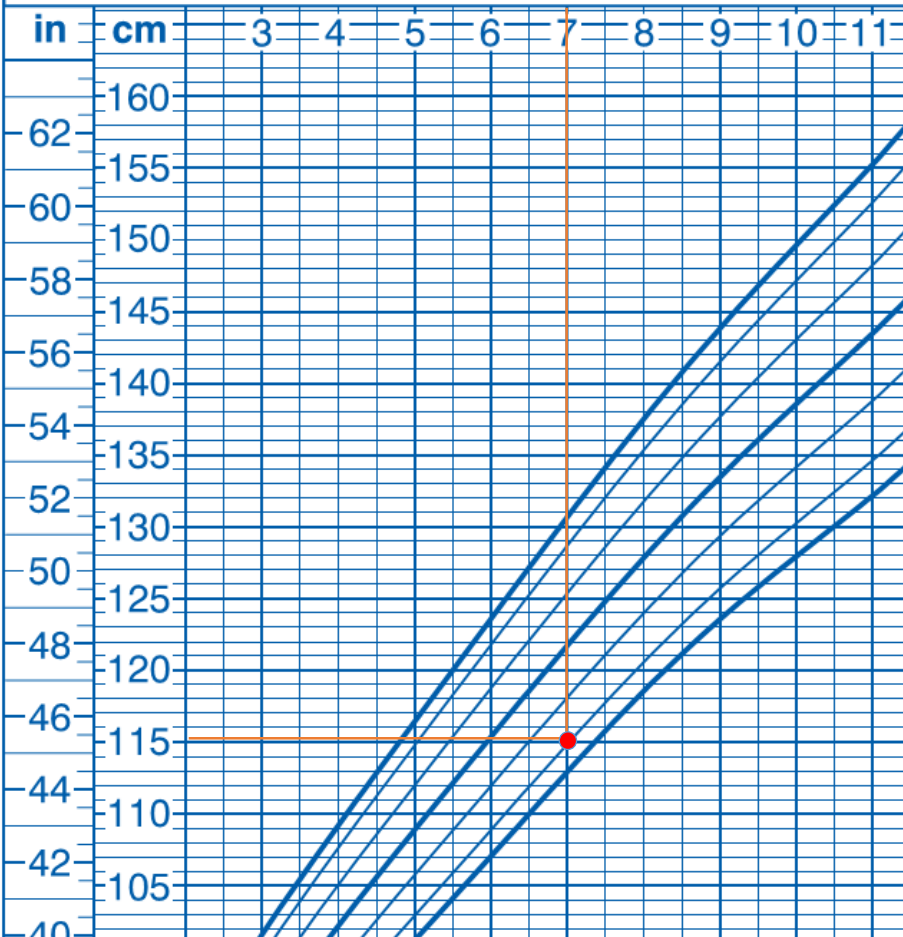
12 13 14 15 16 17 18 19 20

Mother's Stature _____		Father's Stature _____		
Date	Age	Weight	Stature	BMI*
*To Calculate BMI: Weight (kg) ÷ Stature (cm) ÷ Stature (cm) x 10,000 or Weight (lb) ÷ Stature (in) ÷ Stature (in) x 703				

AGE (YEARS)

cm  
in  
STATURE

STATURE



STATURE

W