Deskriptiver Statistik

Datenzusammenfassung
Datendarstellung

Einsicht zu bekommen

G.Schay

Variabilität

Es gibt eine eingebaute unsicherheit und Variabilität in der Natur

das erfordert ein etwas anderes Denken

Statistik beschreibt hauptsächlich zufällige Massenerscheinungen

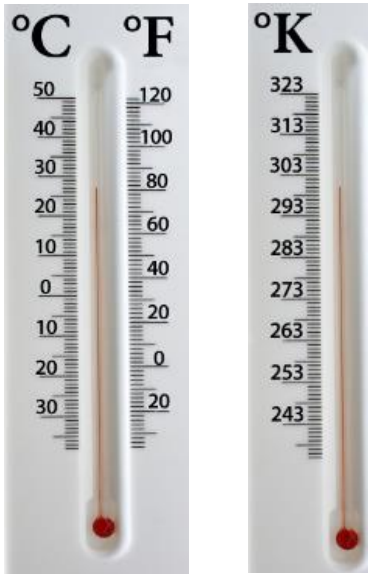
- *Datensammelung (Stichprobe)*
  - *Datenorganisierung*

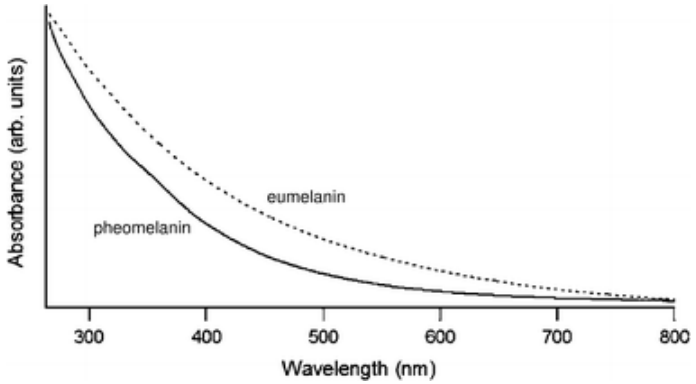
Descriptiver Statistik

- *Datenanalyse*
- *Konklusion,Entscheidung*

Induktiver Statistik

Merkmal, Beobachtung





Merkmaltype

qualitativ
kategorisch

quantitativ
numerisch

intervallskala

verhältnisskala

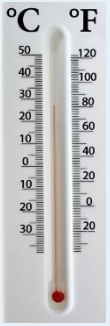
= ≠

= ≠
< >

= ≠
< >
+ −

= ≠
< >
+ −

absoltuer
Nullpunkt



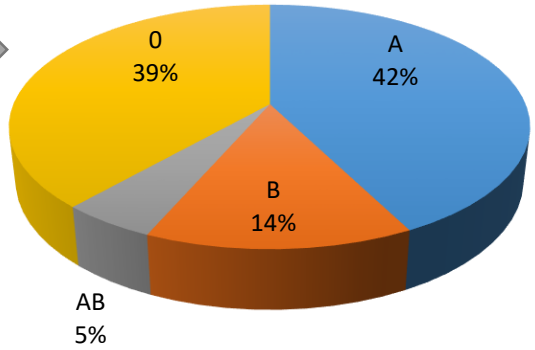
Nominal: wir haben Namen

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

grafische Darstellung

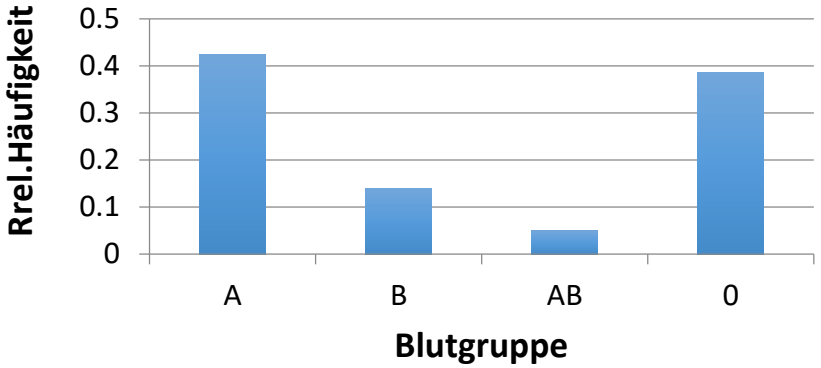
Relative Häufigkeiten



Häufigkeitsverteilung

blood group	(absolute) frequency	relative frequency
A	85	0.425
B	28	0.14
AB	10	0.05
0	77	0.385
Σ	200	1

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

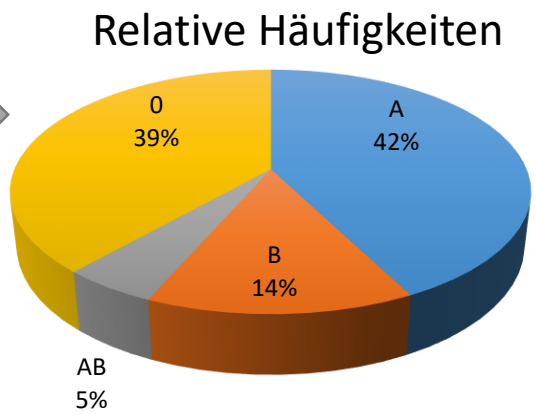
„Mitte?"

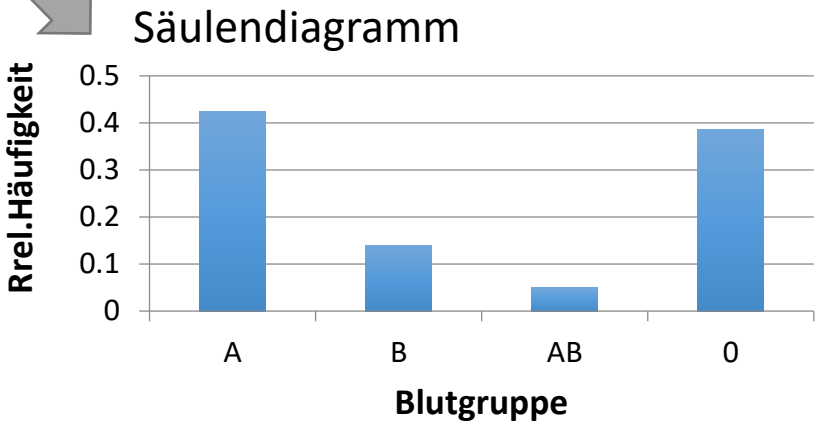
es gibt keine feste Anordnung oder Reihenfolge

häufigster Wert = Modalwert (Modus)

hier: „A"

grafische Darstellung

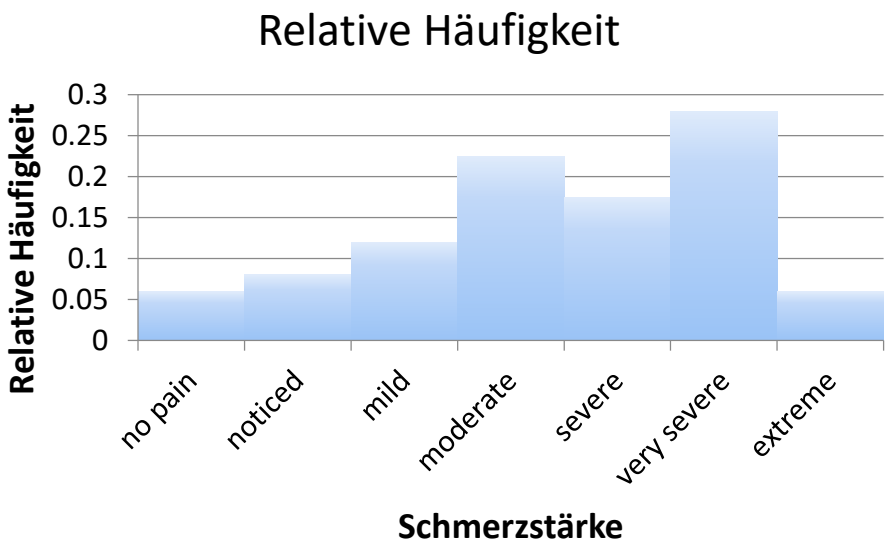




ordinale Merkmale

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
Σ	1	



„Mitte"

Modalwert ist hier (und immer) benutzbar.

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

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

Ordinale Variablen

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
Σ	

Grafisch

wir suchen den X wozu 0.5 als relative kumulierte Häufigkeit gehört

quantitative Merkmale

Häufigkeitsverteilungen

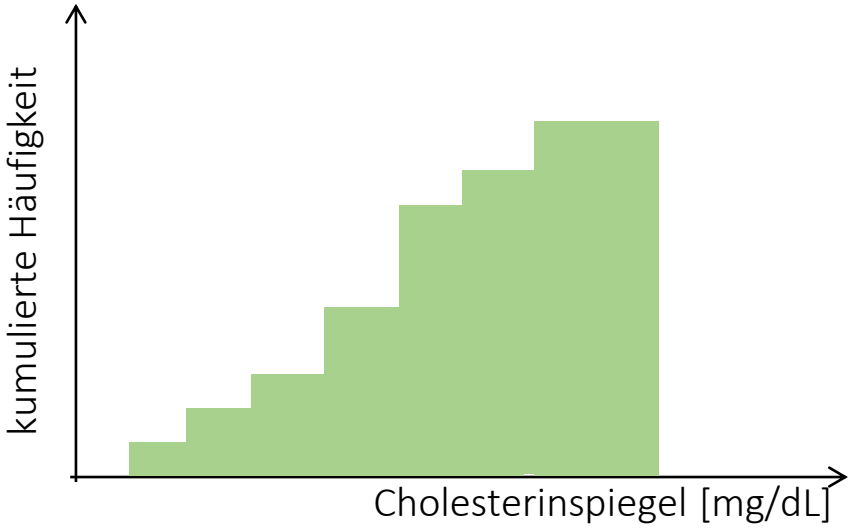
frequency distributions (differential discrimination functions)

Klassen bins (classes, intervals)	abs.H. (absolute) frequency (FREQUENCY) HäUFIGKEIT	rel.H. relative frequency	abs.H.dichte (absolute) frequency density	rel.H.dichte relative frequency density
x ≤ 100	0			
100 < x ≤ 110	0	0	0	0
110 < x ≤ 120	2	0,01	0,2	0,001
120 < x ≤ 130	5	0,025	0,5	0,0025
130 < x ≤ 140	22	0,11	2,2	0,011
140 < x ≤ 150	31	0,155	3,1	0,0155
150 < x ≤ 160	48	0,24	4,8	0,024

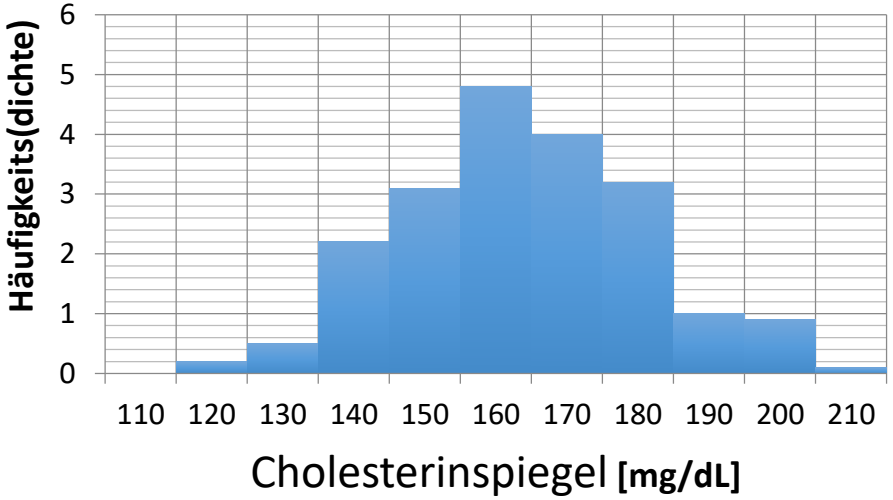
es gibt informationsverlust mit einem Säulendiagramm

Klassenbreite:
praktische Hinsichte (wie sieht gut aus)
Statistisch?
-> Kumuliert is eindeutig!

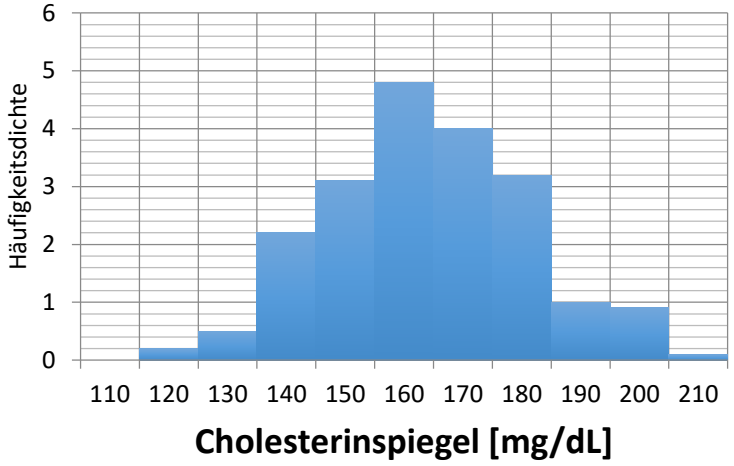
Grafisch

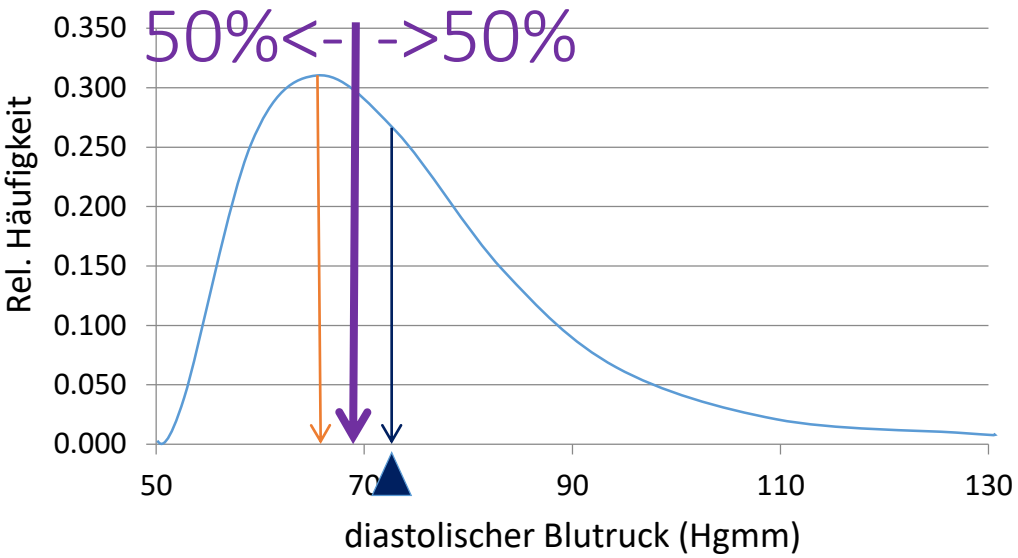


absolute Häufigkeitsverteilung



quantitative Merkmale





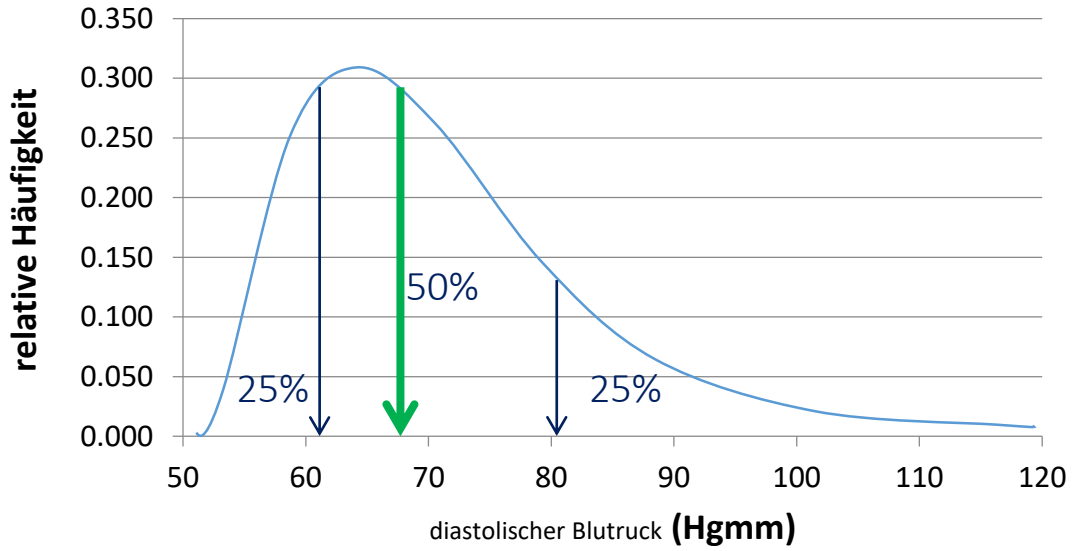
Lageparameter

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

Mittelwert (arithmetisches Mittel, Durchschnitt) mean

$$\bar{x} = x^* = \frac{\sum x_i}{n}$$

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*p Beobachtungen sind kleiner, und maximal N*(1-p) sing größer)

Mittelwert (arithmetisches Mittel, Durchschnitt)
mean

$$\bar{x} = x^* = \frac{\sum x_i}{n}$$

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 entdecken

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

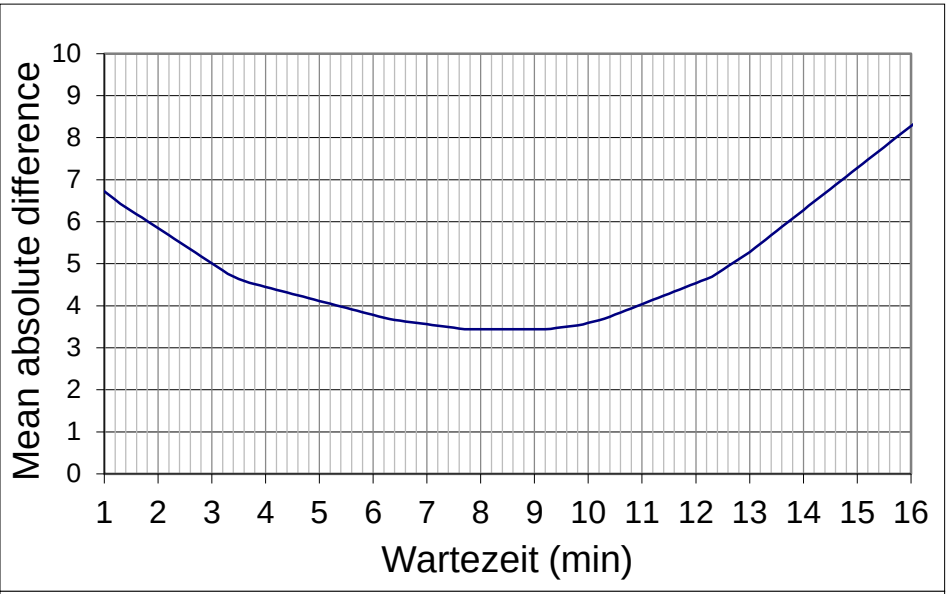
Ergänzungsmaterial

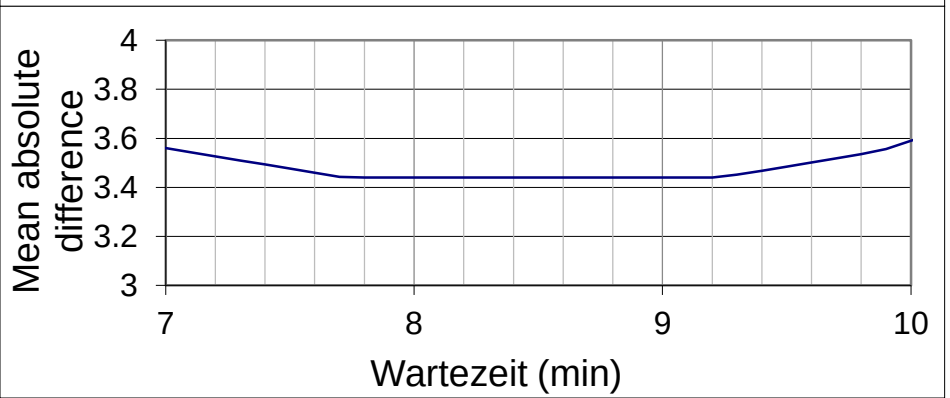
$\frac{1}{n}\sum\left|x_i - x^*\right|$ ist minimal wenn:

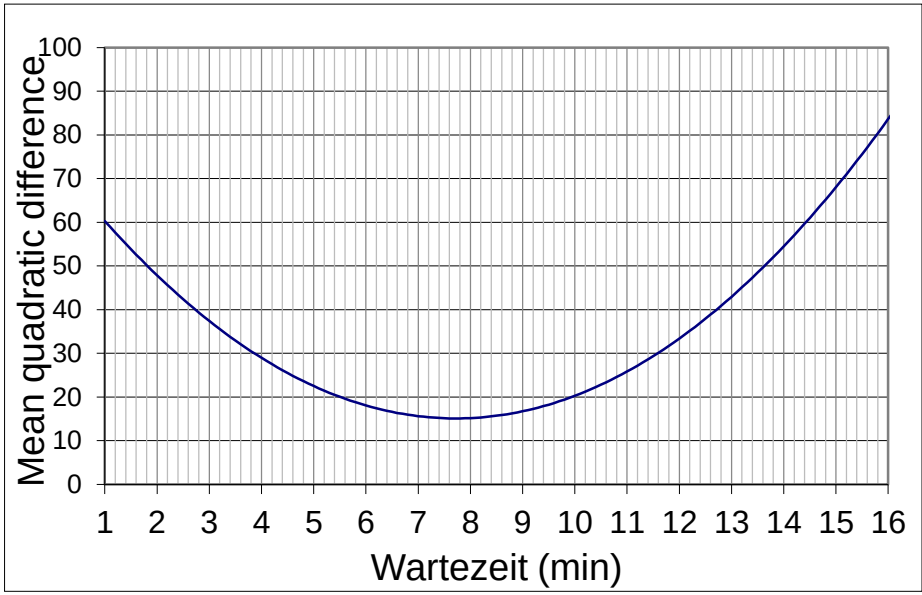
$x^* = Median$

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

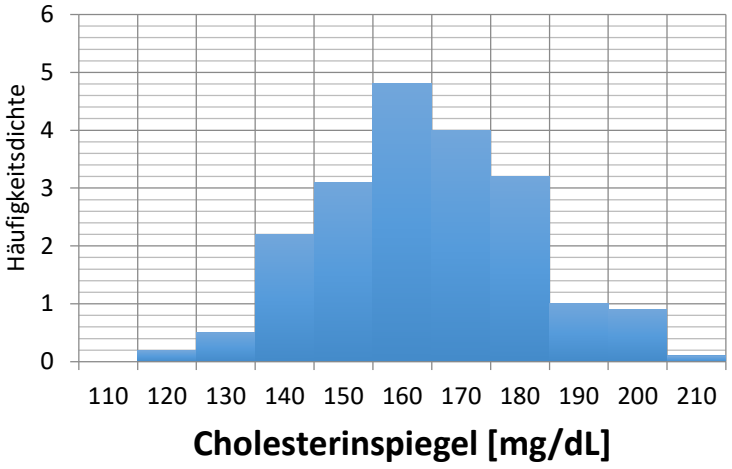
$x^* = Mittelwert$

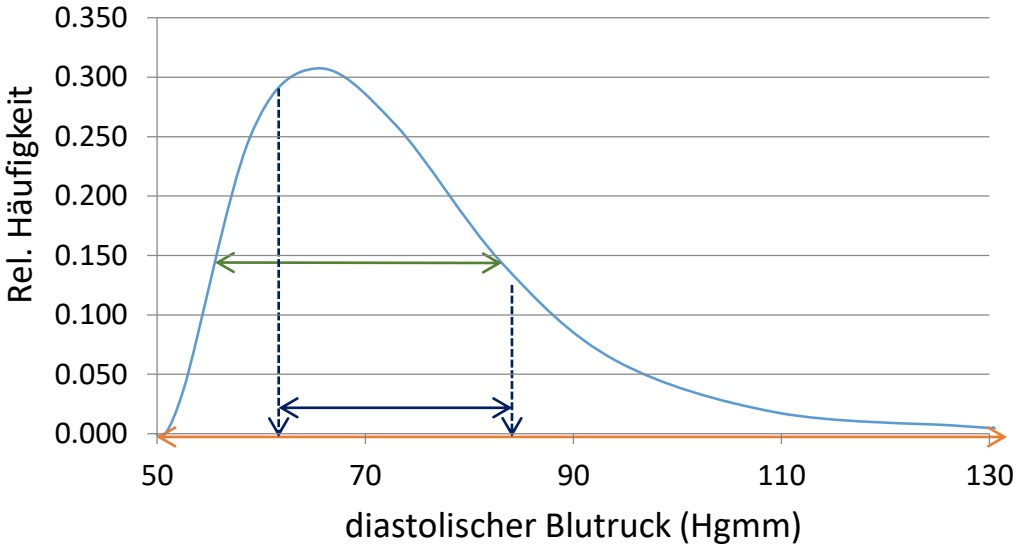






Breiteparameter

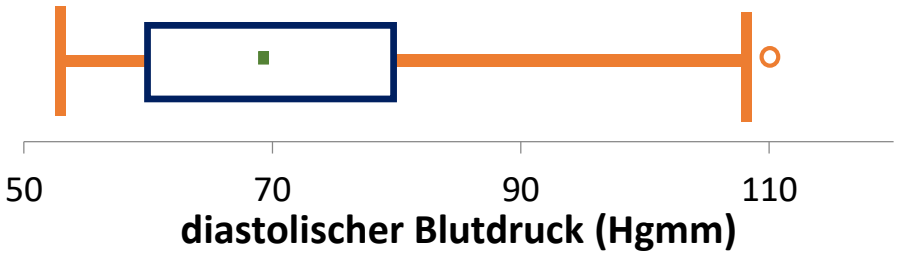




Breite-, Streuungs-parameter

- *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*

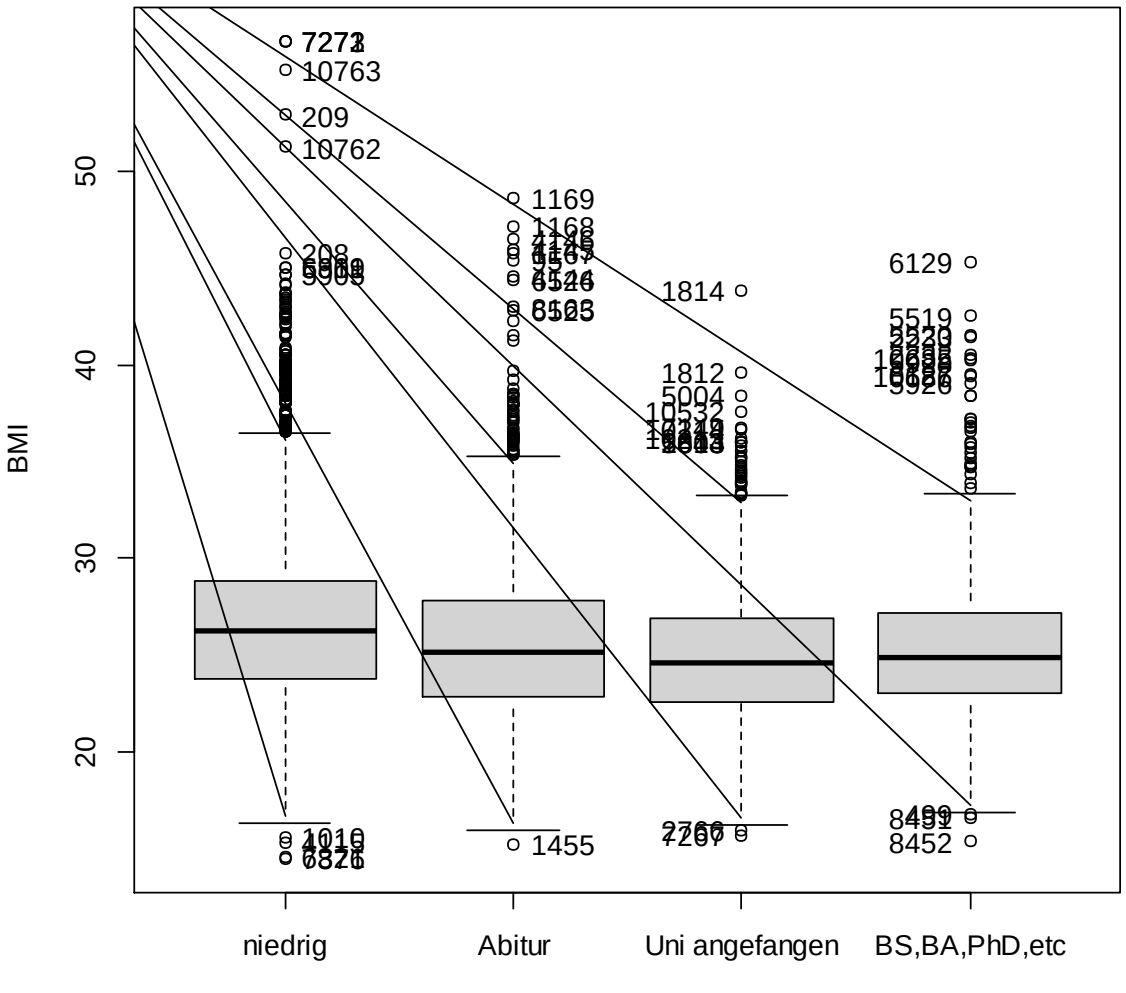
Kästchen: 2*Standardabweichung oder IQR

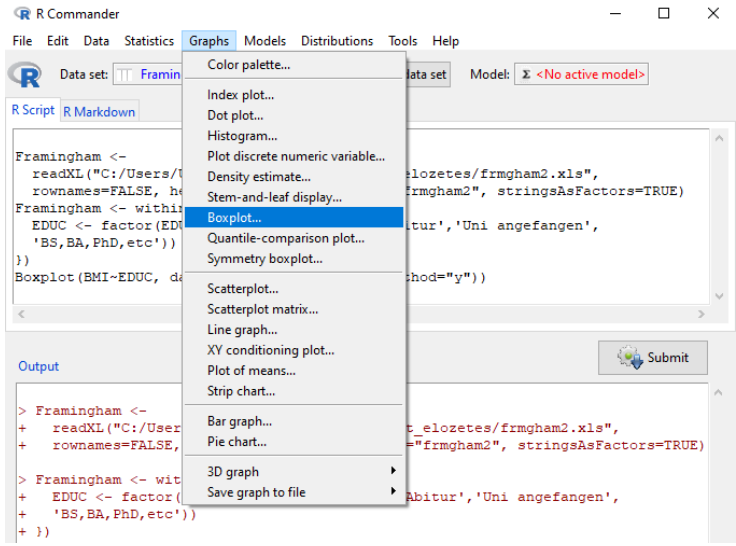
Schnurrhaare (whisker): 3*SD ; minimum und maximum ; 0.05 and 0.95 Quantile, *1.5*IQR*...

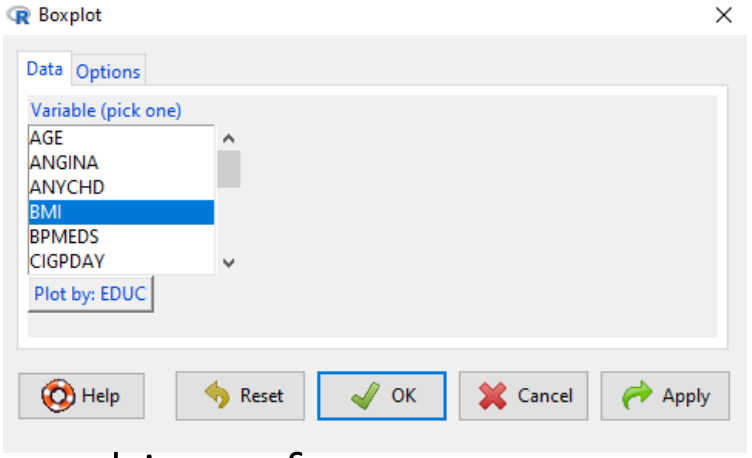
draussen: ***Ausreißer***

abgeschnittener Mittelwert: Mittelwert neugerechnet ohne Ausreißer.

Framingham Datenbank, EDUC geändert zu „Faktoren” mit Namen







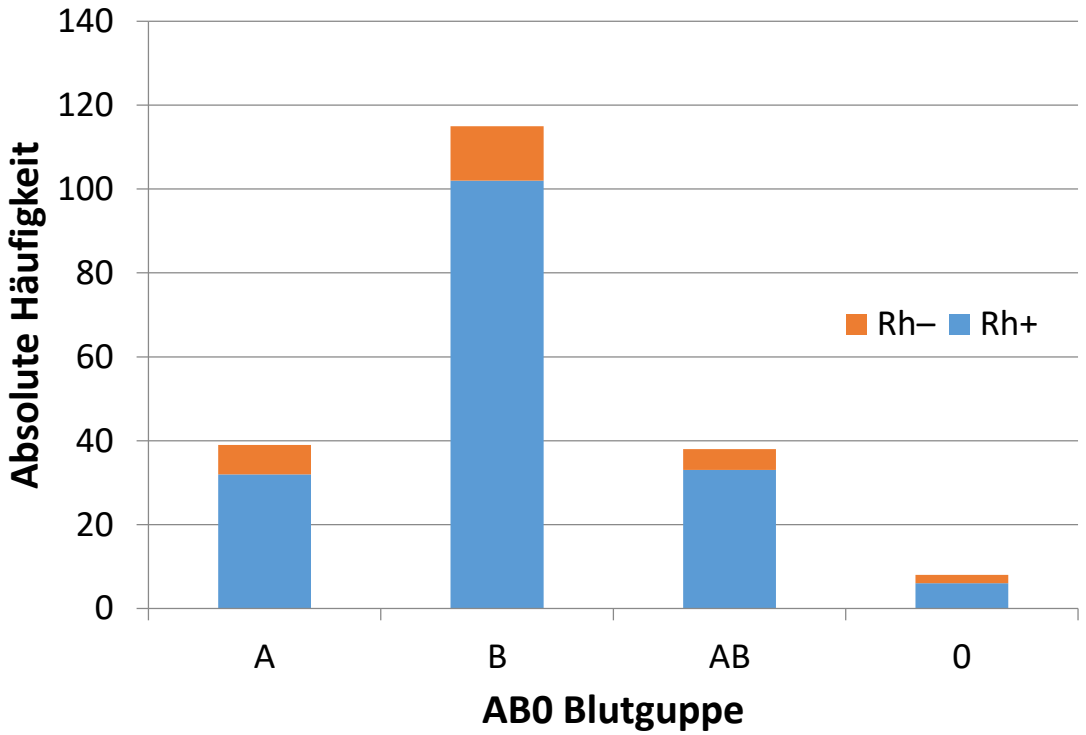
Boxplot (Kastendiagramm): Ausreißer sieht man hier sofort, „wirkliche” unterschiede sind auch enteckbar.

Qualitative Beschreibung mehrerer gemeinsamen Beobachtungen

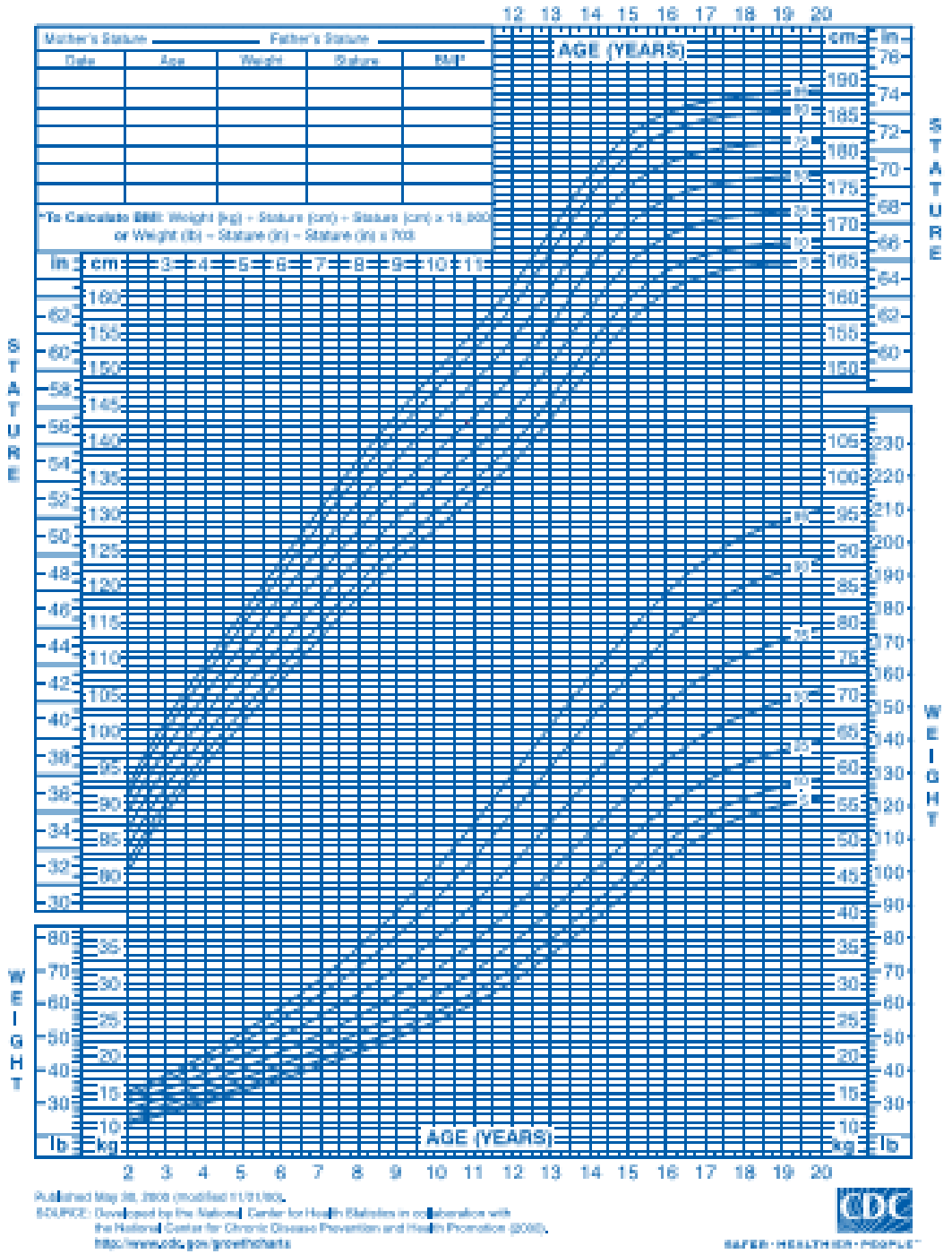
Kontingenztafel

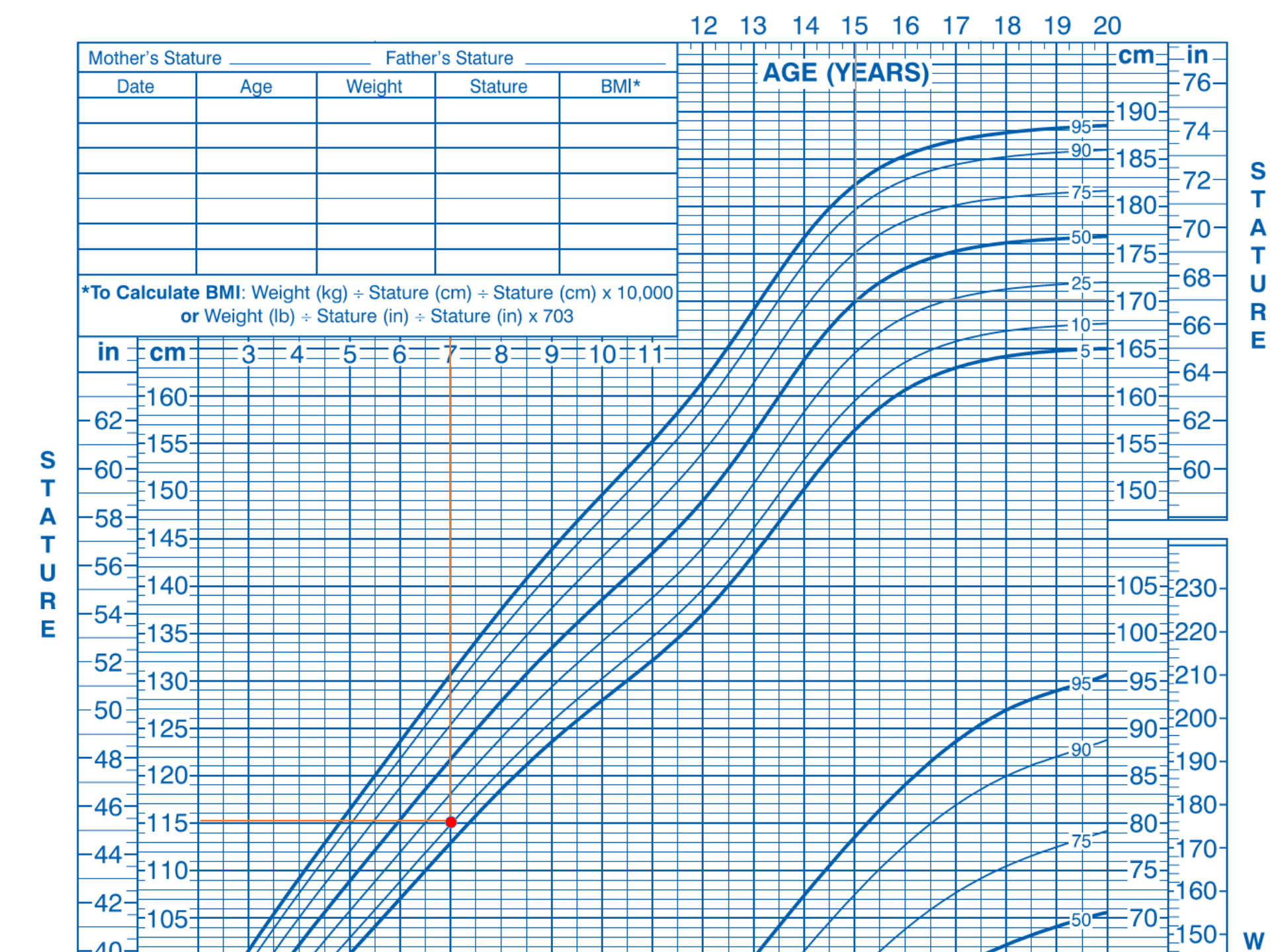
	A	B	AB	0	Σ
Rh+	32	102	33	6	173
Rh–	7	13	5	2	27
Σ	39	115	38	8	200

gestapelter Säulendiagramm



Perzentilkurven




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)
12 13 14 15 16 17 18 19 20
STATURE
in
cm
95
90
75
50
25
10
5
W