Christian-Albrechts-Universität zu Kiel

04_deskriptive_statistik

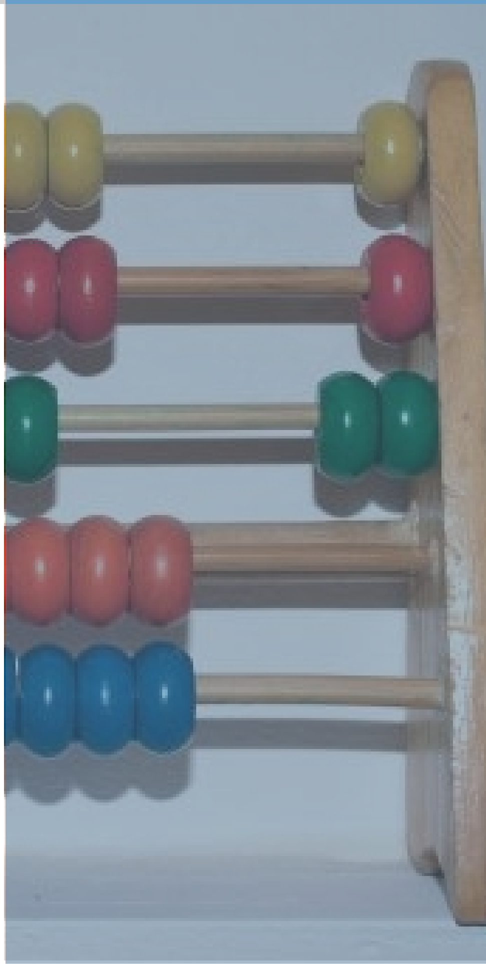
Lage- und Verteilungsmaße

Laden der Daten für weitere Schritte

Einlesen der Daten der Kursteilnehmer:

```
> setwd("--ihr R-Verzeichnis--")
> laender<-read.csv2("laenderdaten.csv")

> laender[1:3,]
                Name Einwohnerzahl Fläche.in.km.                                    Amtssprache        BIP
1 Königreich Dänemark       5732173     2244490.0                                        Dänisch 3.3320e+11
2         New Zealand       4445000      269652.0 Englisch, Maori, neuseeländische Gebärdensprache 1.6181e+11
3            Schweden       9644864      438575.8                                      Schwedisch 5.3820e+11
  Weltrang.nach.BIP Weltrang.CPI Einlieferer kontinent
1                32            1      breske    Europa
2                56            1      breske      <NA>
3                21            1      breske    Europa
```



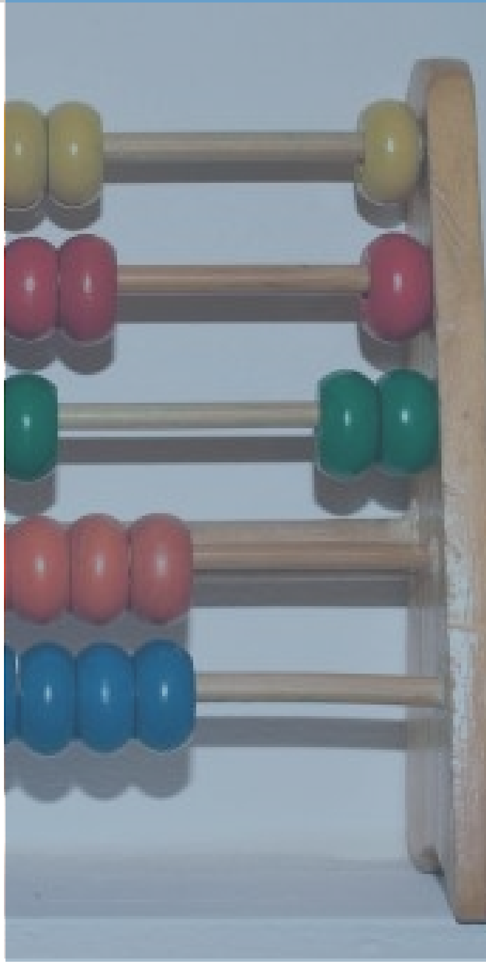
Deskriptive Statistik

Summarische Darstellung einer Menge von beobachteten Daten
Die Verteilung der Daten innerhalb der Stichprobe wird wiedergegeben

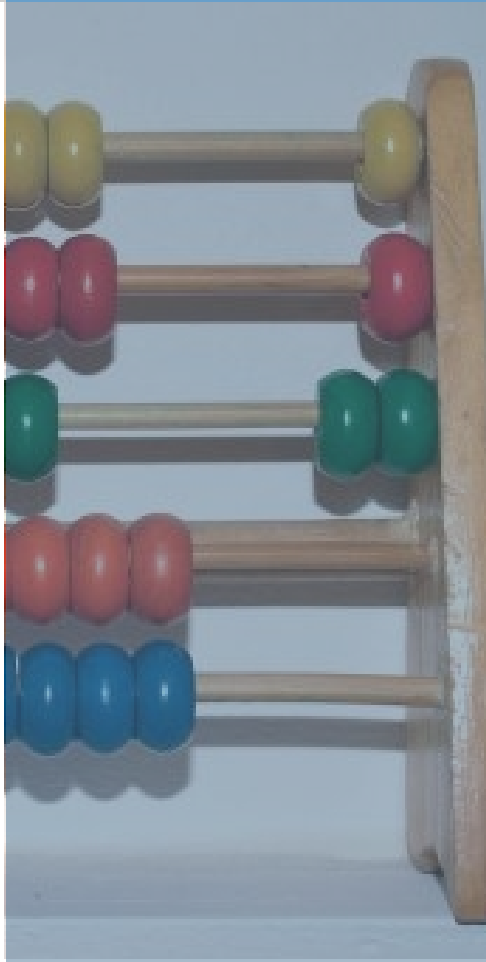
Darstellungsarten

Tabellarisch – Kontingenztabelle
Graphisch – Diagramme
Numerisch – Mit Hilfe von Kennwerten für die Verteilung

Deskriptive Statistik macht (eigentlich) keine Aussagen über die Grundgesamtheit, sondern beschreibt nur die Stichprobe! (im Unterschied hierzu Inferenzstatistik)



Aspekte von Verteilungen



Tendenz zur Mitte (zentrale Tendenz):
Gibt an, wo in der Spannweite der Werte die Mitte liegt

Arithm. Mittel, Median, Modus

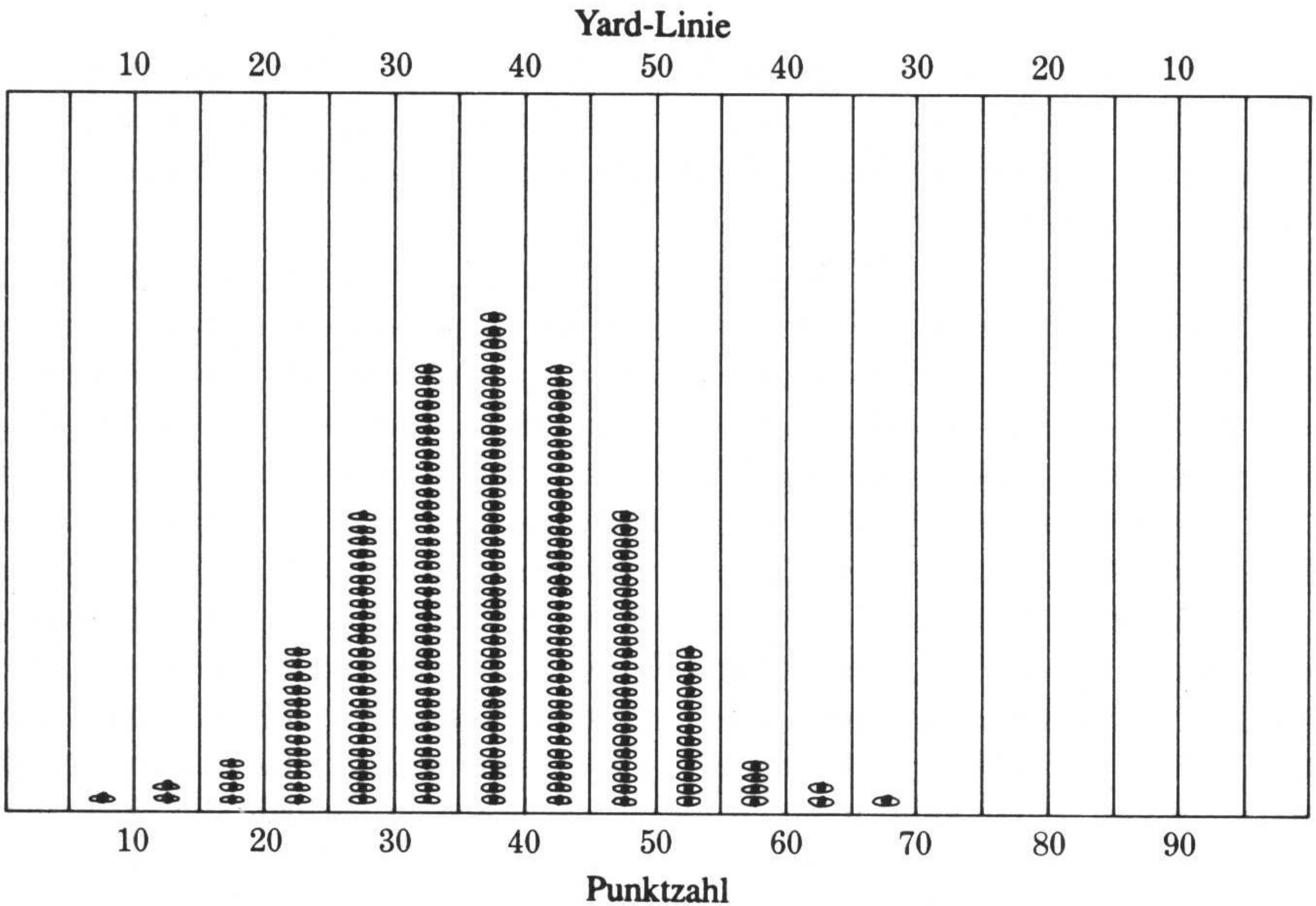
Streuung:
Gibt an, wie breit die Werte streuen

Variationsbreite, Varianz, Standardabweichung, Variationskoeffizient

Form:
Form der Verteilungskurve

Symmetrisch/Asymmetrisch

Schiefe und Kurtosis(Wölbung)



Studenten, die sich nach ihren Testergebnissen in Reihen auf einem Footballfeld aufgestellt haben – eine Häufigkeitsverteilung.

Quelle: Phillips 1997

Tendenz zur Mitte [1]

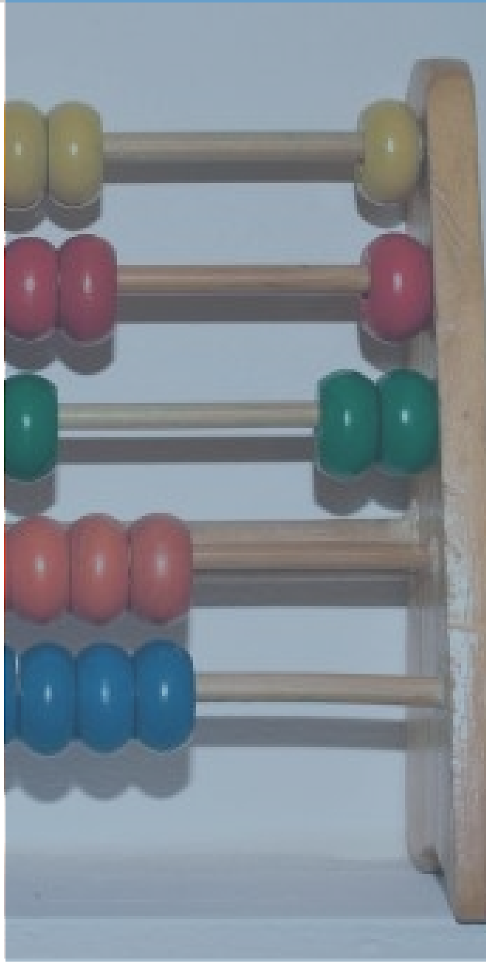
Arithmetisches Mittel

Der Klassiker, auch Durchschnitt oder Mittelwert genannt. Geht für metrische Daten (intervall oder verhältnis)

Summe der Werte/Anzahl der Werte, oder $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

R:

```
> sum(laender$Fläche)/length(laender$Fläche)
[1] 943844
> mean(laender$Fläche)
[1] 943844
```



Tendenz zur Mitte [2]

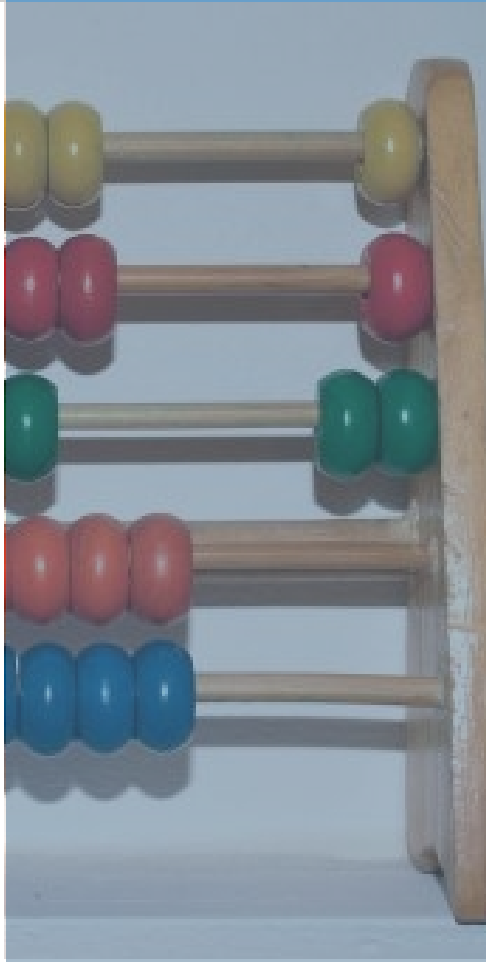
Median
Läßt sich für metrische und ordinale Variablen bestimmen.

Bei ungerader Anzahl: der mittlere Wert einer sortierten Reihe

```
1 2 3 4 5 6 7
      |
R:
> median(c(1,2,3,4,5,6,7))
[1] 4
```

Bei gerader Anzahl: das Arithm. Mittel der mittleren Werte einer sortierten Reihe

```
1 2 3 4 5 6 7 8
       |
R:
> median(c(1,2,3,4,5,6,7,8))
[1] 4.5
```



Tendenz zur Mitte [3]

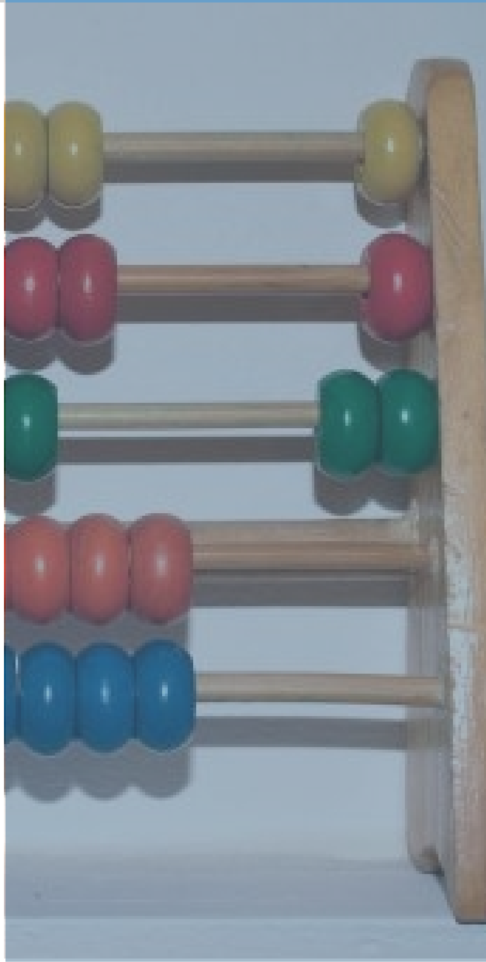
Modus (Modalwert)
Der häufigste Wert einer Datenreihe. Läßt sich für metrische, ordinale und nominale Variablen bestimmen.

Ziege Schaf Ziege Rind Rind Ziege Schwein Ziege

Modus: Ziege

In R:

```
> which.max(table(c("Ziege", "Schaf", "Ziege", "Rind",
"Rind", "Ziege", "Schwein", "Ziege")))
Ziege
    4
```

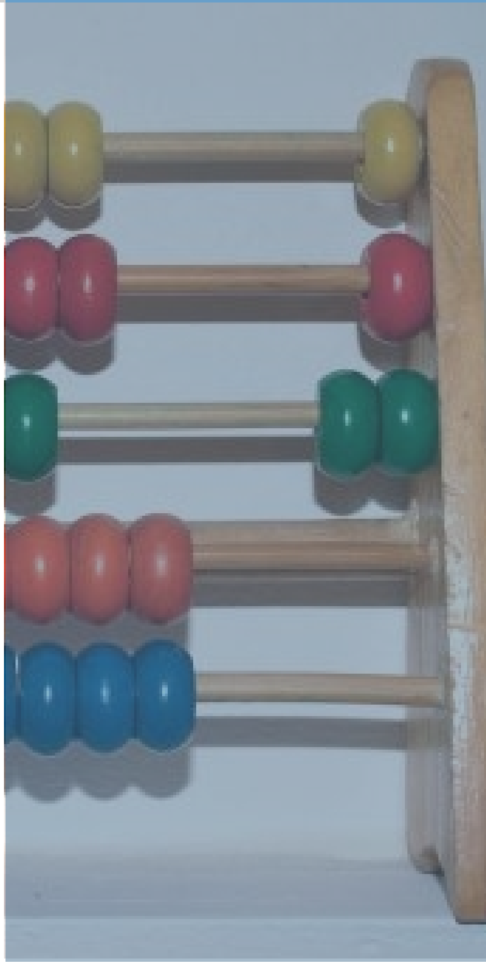


Christian-Albrechts-Universität zu Kiel

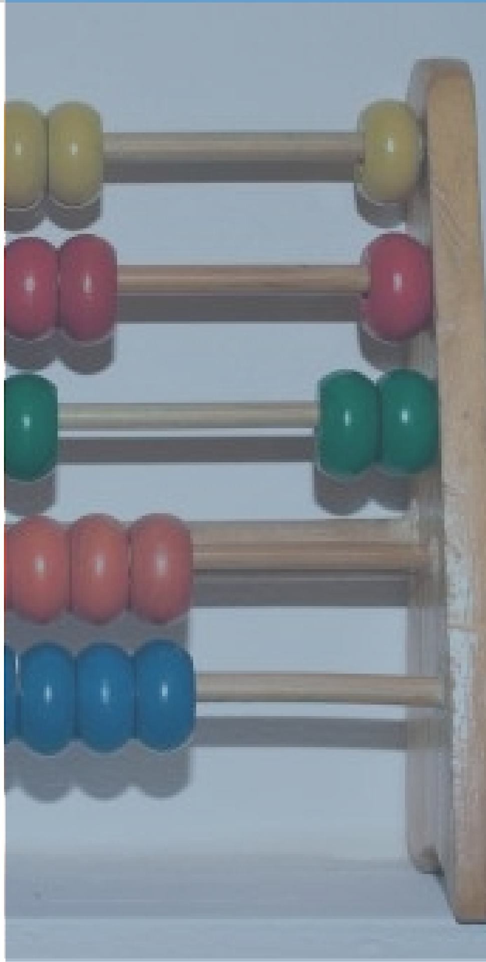
Tendenz zur Mitte [4]

Merkmal ist...		
nominal-skaliert	ordinal-skaliert	intervall-skaliert+
Modus	Modus	Modus
-	Median	Median
-	-	Arith. Mittel

Nach: Dolić 2004



Tendenz zur Mitte [5]



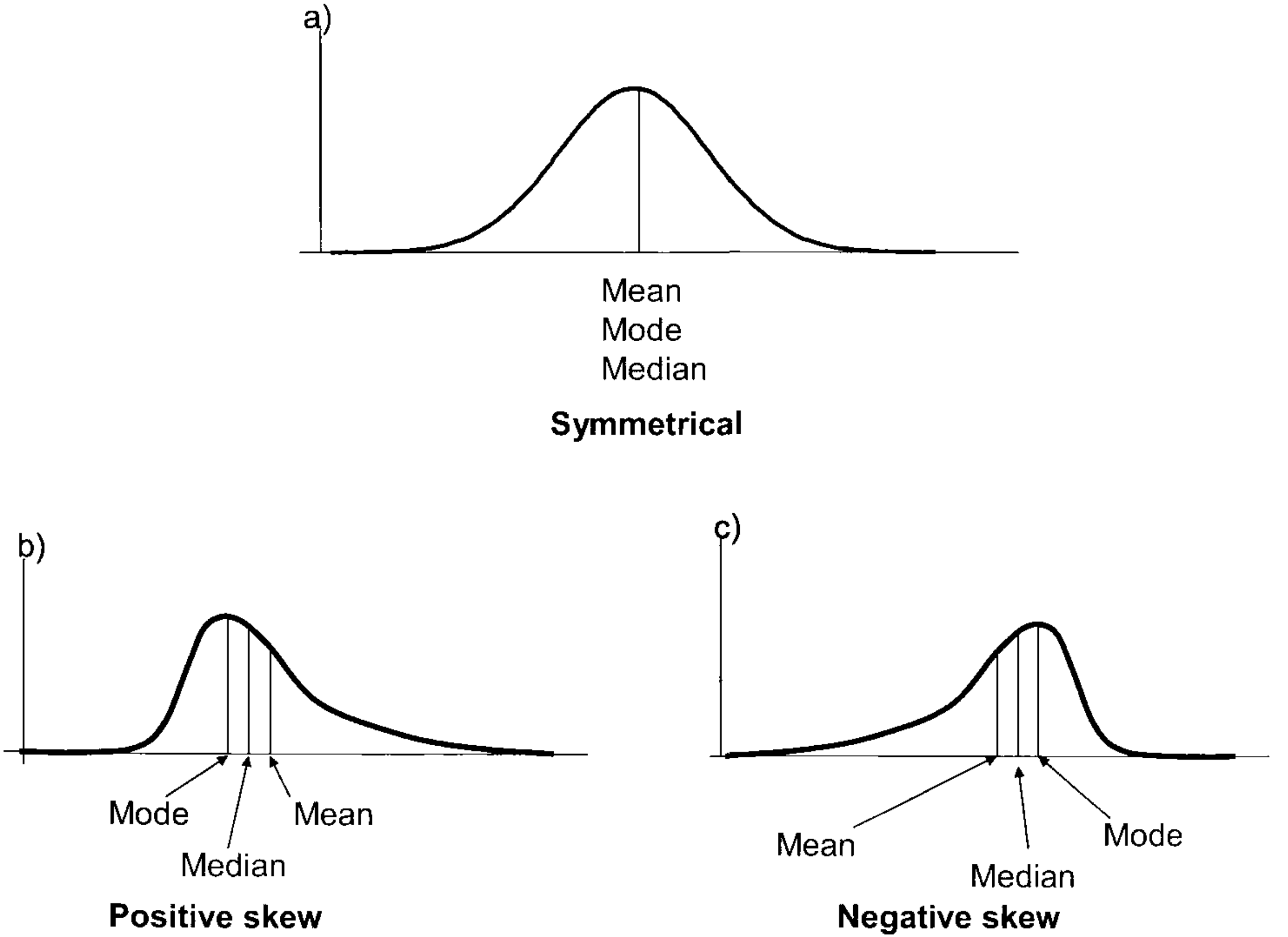
Vergleich der Mittelwerte

Anfälligkeit für Ausreißer: Der Mittelwert ist sehr anfällig für Ausreißer, der Median deutlich weniger, der Modus kaum.

```
> test<-c(1,2,2,3,3,3,4,4,5,5,6,7,8,8,8,9,120)
> mean(test)
[1] 11.64706
> median(test)
[1] 5
> which.max(table(test))
3
3
```

Der Modus eignet sich kaum für die Beschreibung von metrischen oder nominalen Daten, nur dann, wenn eine einigermaßen symmetrische Verteilung vorliegt.

```
> which.max(table(c(1,2,2,3,3,3,4,4,4,4,5,5,5,6,6,7)))
4
4
```


a)
Mean
Mode
Median
Symmetrical
b)
Mode
Mean
Median
Positive skew
c)
Mean
Mode
Median
Negative skew

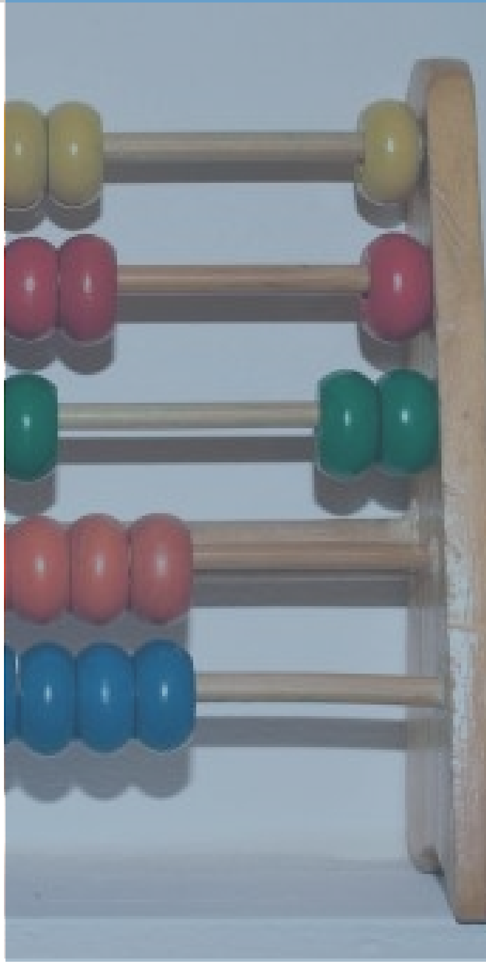
Tendenz zur Mitte Aufgabe

Beschreiben Sie die Mittelwerte

Zu analysieren sind die Messungen der Breite an Tassen aus dem Gräberfeld Walternienburg in cm (Müller 2001, 534; Auswahl):

```
> tassen<-read.csv2("tassen.csv",row.names=1)
> tassen$x
```

Bestimmen Sie Modus, Median und Arith. Mittel und geben Sie an, ob die Schiefe positiv (rechtsschief) oder negativ (linksschief) ist.



Tendenz zur Mitte Aufgabe

Beschreiben Sie die Mittelwerte

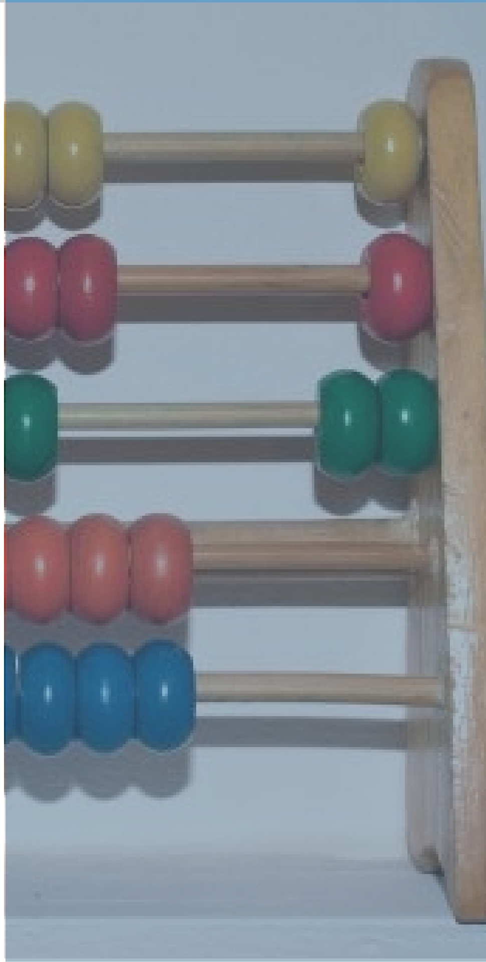
Zu analysieren sind die Messungen der Breite an Tassen aus dem Gräberfeld Walternienburg in cm (Müller 2001, 534; Auswahl):

```
> tassen<-read.csv2("tassen.csv",row.names=1)
> tassen$x
```

Bestimmen Sie Modus, Median und Arith. Mittel und geben Sie an, ob die Schiefe positiv (rechtsschief) oder negativ (linksschief) ist.

```
> mean(tassen$x)
[1] 13.67727
> median(tassen$x)
[1] 12
> which.max(table(tassen$x))
8.1
  3
```

Der Median ist kleiner als das Arithm. Mittel: positiv (rechtsschief).



Grundlegende statistische Verfahren für archäologische Datenanalyse in R

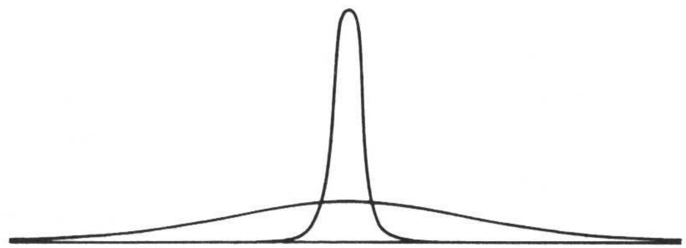
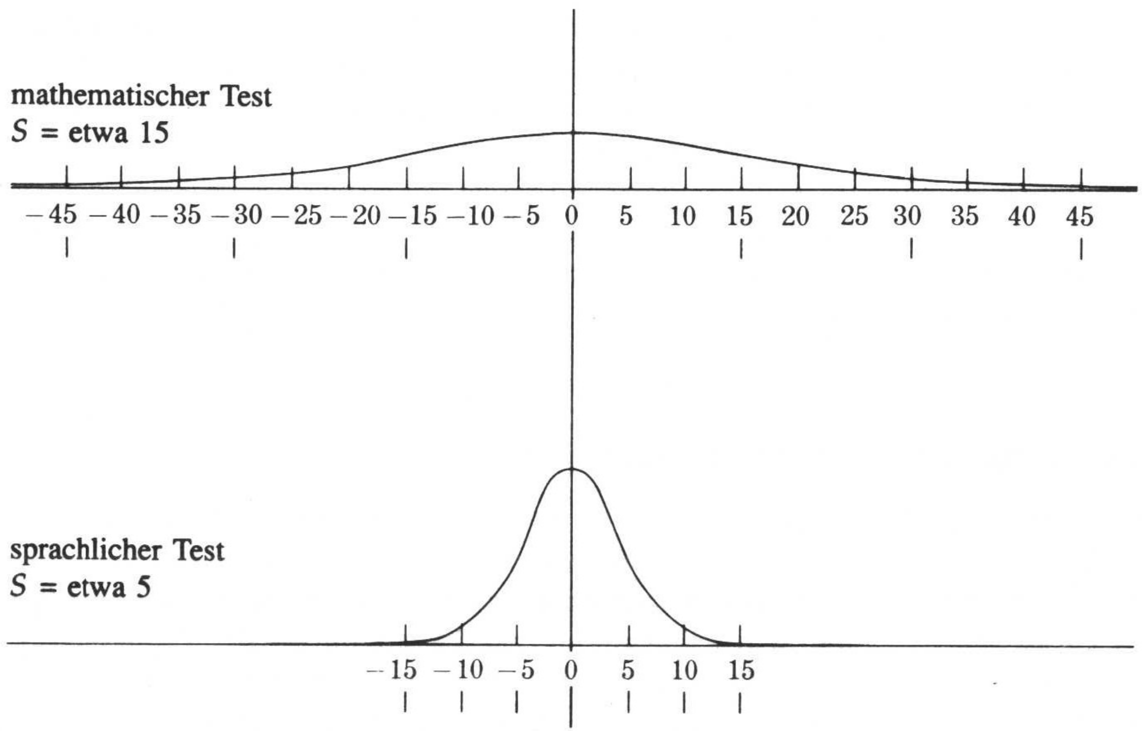


Abb. 4.1 Zwei Verteilungen mit denselben *N*s, aber unterschiedlicher Streuung.



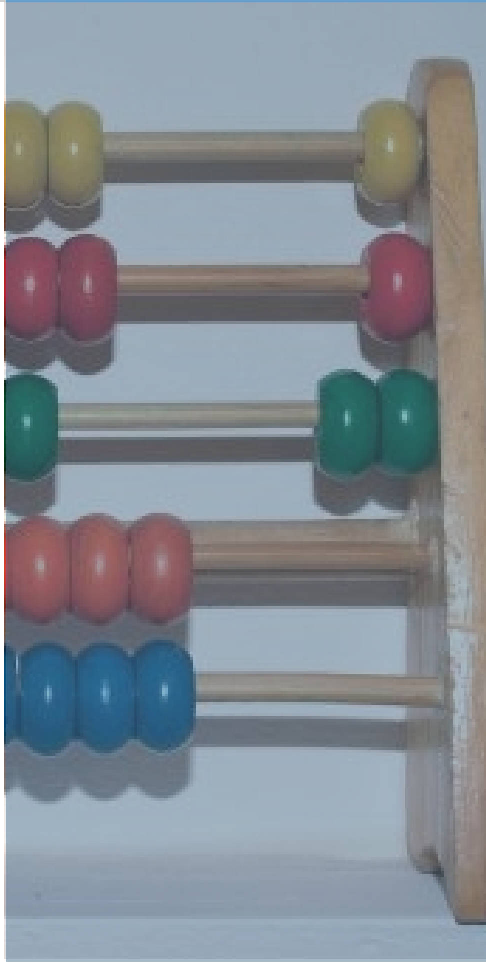
Quelle: Phillips 1997

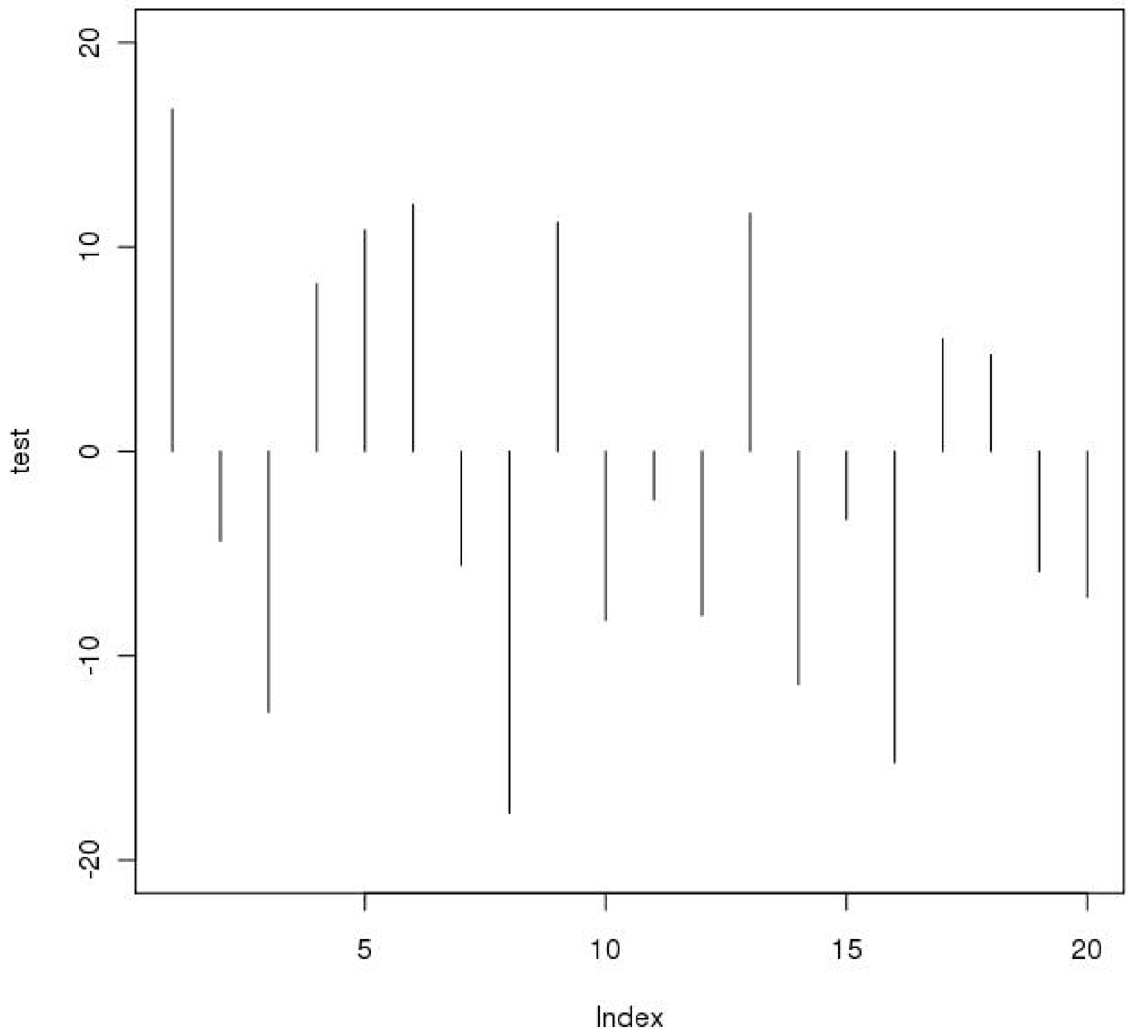
Streuung [1]

Variationsbreite
Einfach die Spannweite der Werte in einer Datenreihe

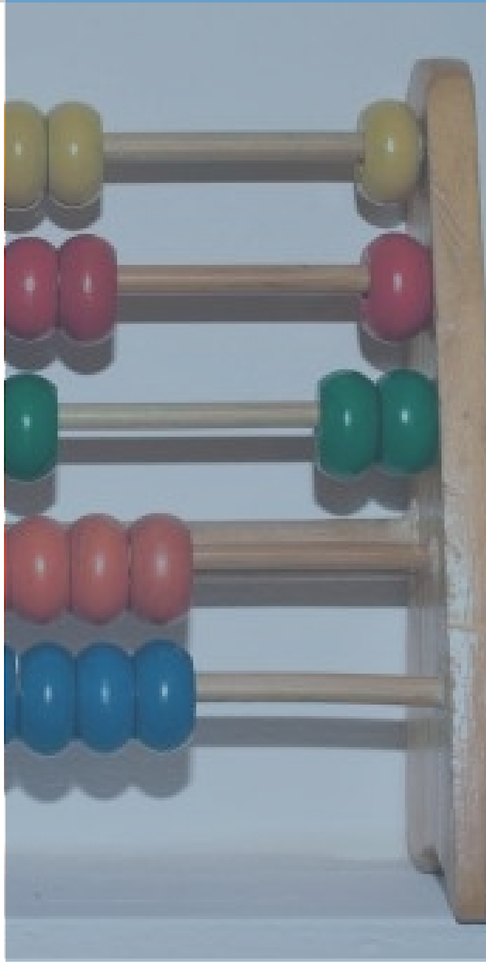
```
> range(laender$Fläche)
[1]    14954 9826675
> range(tassen$x)
[1]  7.5 26.1
```

Da sich das Maß auf die Extremwerte bezieht, ist es logischerweise sehr ausreißeranfällig




20
10
0
-10
-20
test
5
10
15
20
Index

Streuung [2]



(empirische) Varianz

Maß für die Variabilität in den Daten, unabhängiger gegen Ausreißer

Entspricht der Summe der quadrierten Abstände zum Mittelwert durch die Anzahl der Beobachtungen

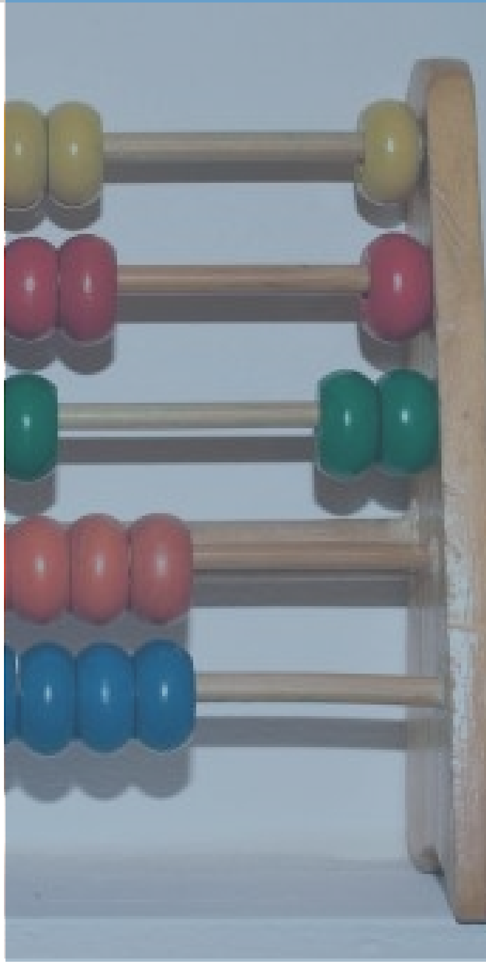
$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

In R per Hand:

```
> sum((tassen$x-mean(tassen$x))^2)/(length(tassen$x)-1)
[1] 31.11136
> var(tassen)
         x
x 31.11136
```

Achtung: es gibt da noch die andere Varianz σ^2 (mit n statt n-1), die ist aber nur für die Grundgesamtheit (die meist nicht bekannt ist), nicht für Stichproben anwendbar.

Streuung [3]



(empirische) Standardabweichung

Varianz hat durch Quadrierung quadrierte Einheiten (mm → mm²)

Um Kennzahl mit ursprünglichen Einheiten vergleichbar zu machen:
Wurzel ziehen: Standardabweichung

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

```
> sqrt(sum((tassen$x-mean(tassen$x))^2)/(length(tassen$x)-1))

> sd(tassen$x)
```

Entspricht sozusagen der durchschnittlichen Abweichung vom Mittelwert

Achtung: es gibt da noch die andere Standardabweichung σ (mit n statt n-1), die ist aber nur für die Grundgesamtheit (die meist nicht bekannt ist), nicht für Stichproben anwendbar.

Streuung [4]

Variations-Koeffizient

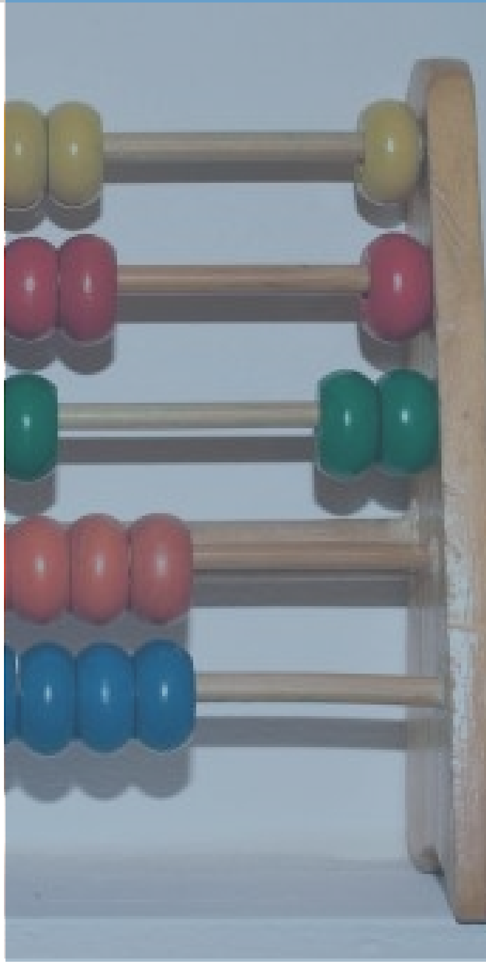
Standardabweichung liegt in der jeweiligen Einheit (z.B. mm) vor

Zum Vergleich zweier Zahlenreihen mit unterschiedlichen Einheiten:
Variantionskoeffizient=Standardabweichung/Mittelwert

Bsp. Variieren Fläche und Einwohnerzahl der Länder etwa gleich stark?

```
> sd(laender$Fläche)/mean(laender$Fläche)
[1] 2.576648
> sd(laender$Einwohnerzahl)/mean(laender$Einwohnerzahl)
[1] 2.479968
```

Einwohnerzahl variieren etwas schwächer als Fläche

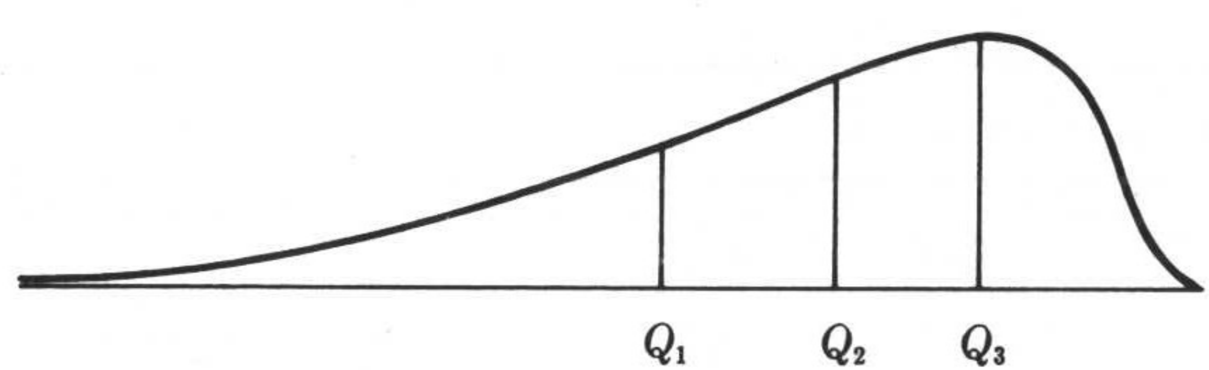


Streuung [5]

Quantile

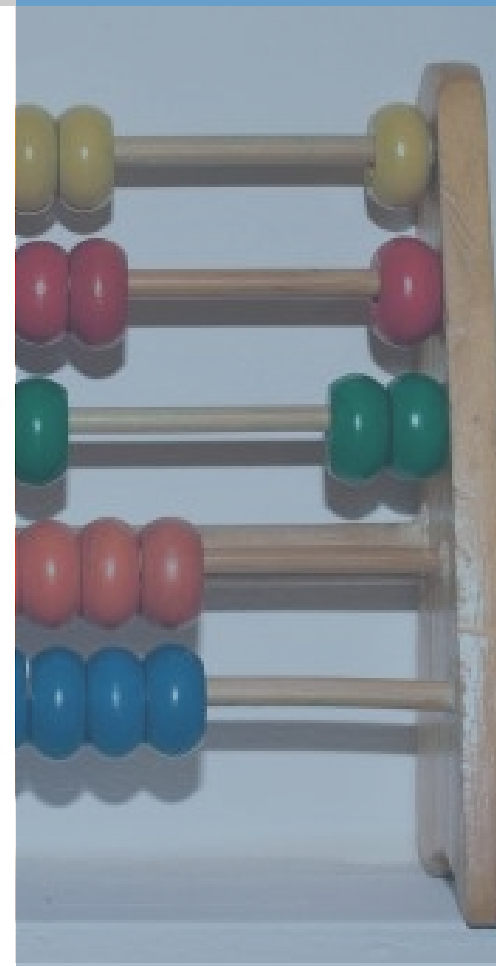
Das 1., 2., 3. und 4. Viertel der Daten (sortiert und durchgezählt) bzw. deren Trennwerte

```
> quantile(tassen$x)
```



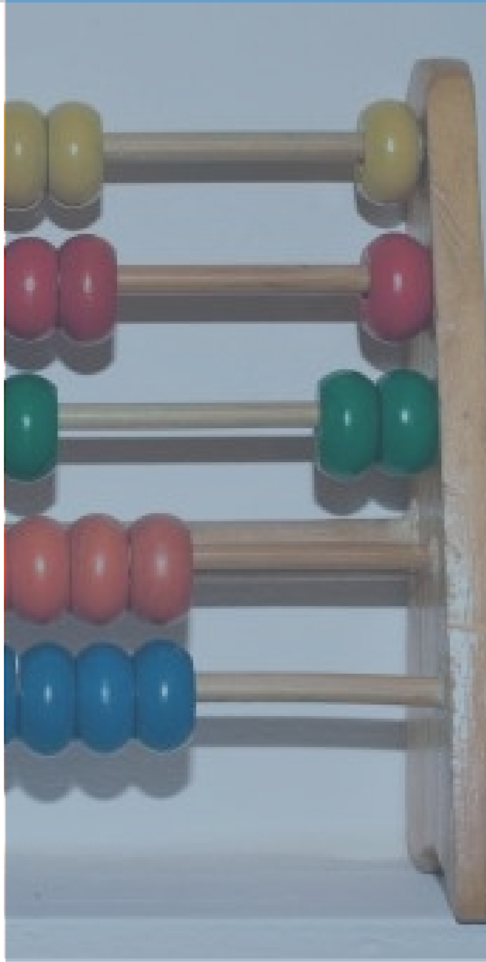
Linksschiefe Verteilung mit einer in Viertel geteilten Fläche.

Quelle: Phillips 1997



C A U
Christian-Albrechts-Universität zu Kiel

Streuung [5]



Quantile

Das 1., 2., 3. und 4. Viertel der Daten (sortiert und durchgezählt) bzw. deren Trennwerte

```
> quantile(tassen$x)
  0%   25%   50%   75% 100%
 7.5   9.0 12.0 18.9 26.1
```

Jetzt neu: Perzentile (Das gleiche für Zehntel)

```
> quantile(tassen$x, probs=seq(0,1,0.1))
   0%    10%    20%    30%    40%    50%    60%    70%    80%    90%   100%
 7.50   8.10   8.52   9.27 10.02 12.00 13.08 18.81 19.38 20.31 26.10
```

Streuungmaß Innerquartilsabstand

```
> IQR(tassen$x)
[1] 9.9
```

Unempfindlicher gegen Ausreißer als Standardabweichung, dafür geht Information verloren

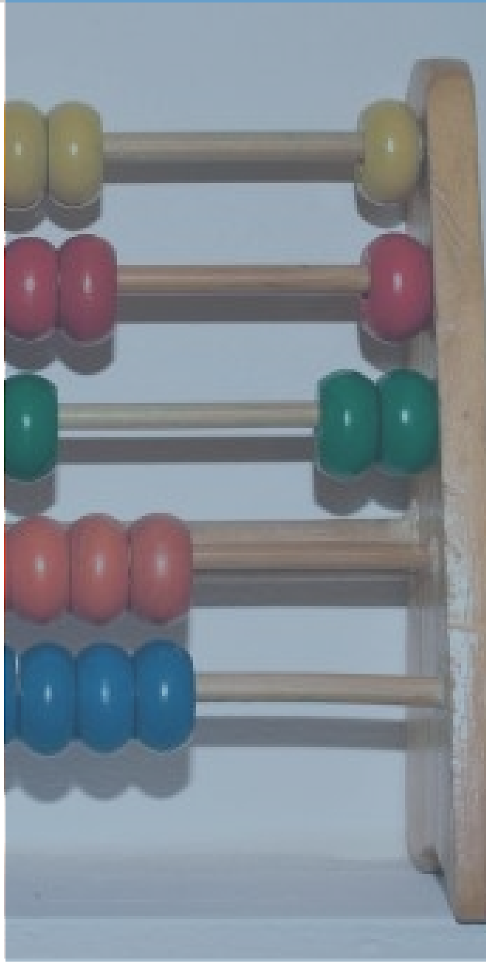
Streuung Aufgabe

Bestimmen Sie die Streuung der Daten

Zu analysieren sind die Größen der von versch. Megalithgräbern aus sichtbaren Flächen (Demnick 2009):

```
> altmark<-read.csv2("altmark_denis2.csv",row.names=1)
> altmark$sichtflaeche
```

Finden Sie heraus, in welcher der Regionen die Gräber eine einheitlichere Sichtfläche haben.



Streuung Aufgabe

Bestimmen Sie die Streuung der Daten

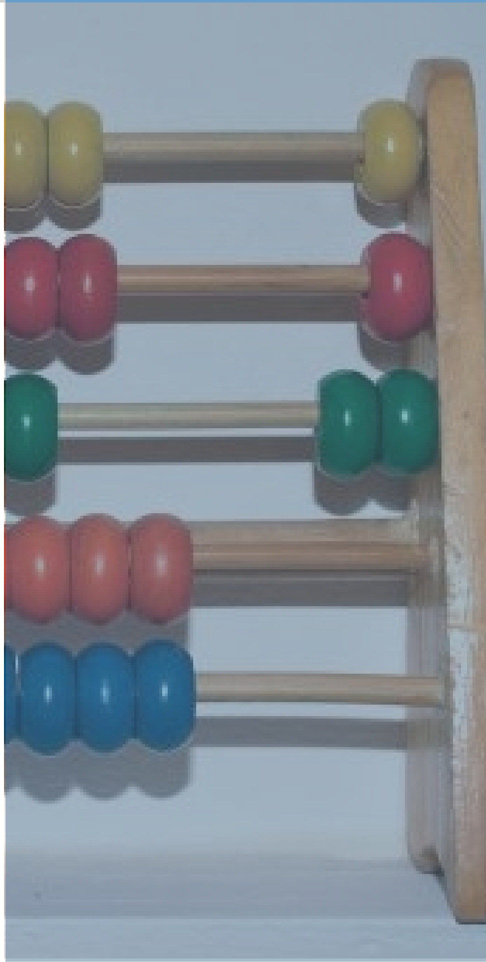
Zu analysieren sind die Größen der von versch. Megalithgräbern aus sichtbaren Flächen (Demnick 2009):

```
> altmark<-read.csv2("altmark_denis2.csv",row.names=1)
> altmark$sichtflaeche
```

Bestimmen Sie für jede der Regionen die Standardabweichung.

```
> sd(altmark[altmark$region=="Mitte",1])
[1] 60.56687
> sd(altmark[altmark$region=="Ost",1])
[1] 51.46048
> sd(altmark[altmark$region=="West",1])
[1] 28.73535
```

Die Standardabweichung ist für die Region West am kleinsten, die Sichtflächen sind hier am einheitlichsten



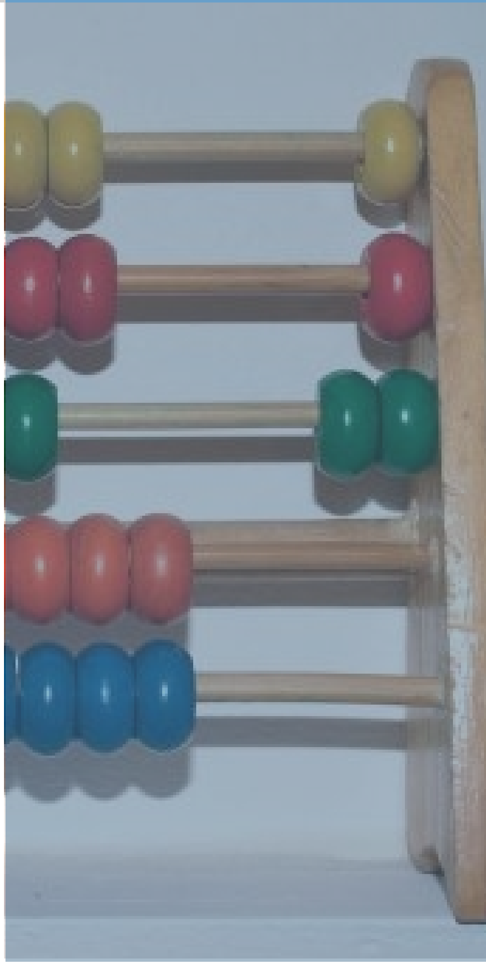
Form der Verteilung [1]

Wichtige Parameter

Anzahl der Gipfel der Verteilung: unimodal, bimodal, multimodal

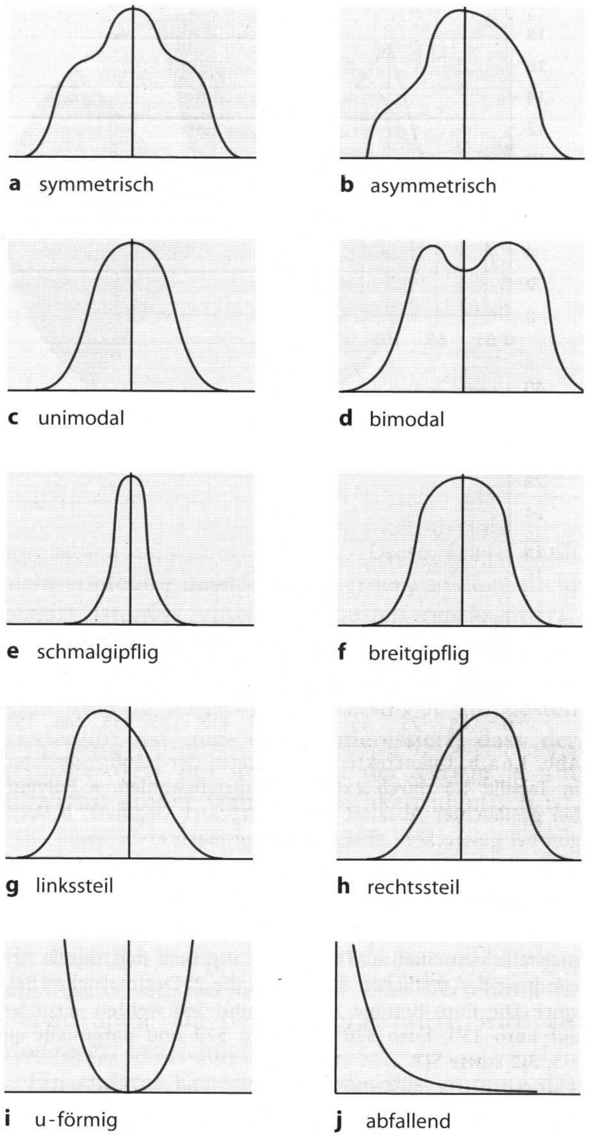
Schiefe der Verteilung: Rechtsschief, Linksschief

Kurtosis (Wölbung): flach, mittel, steil

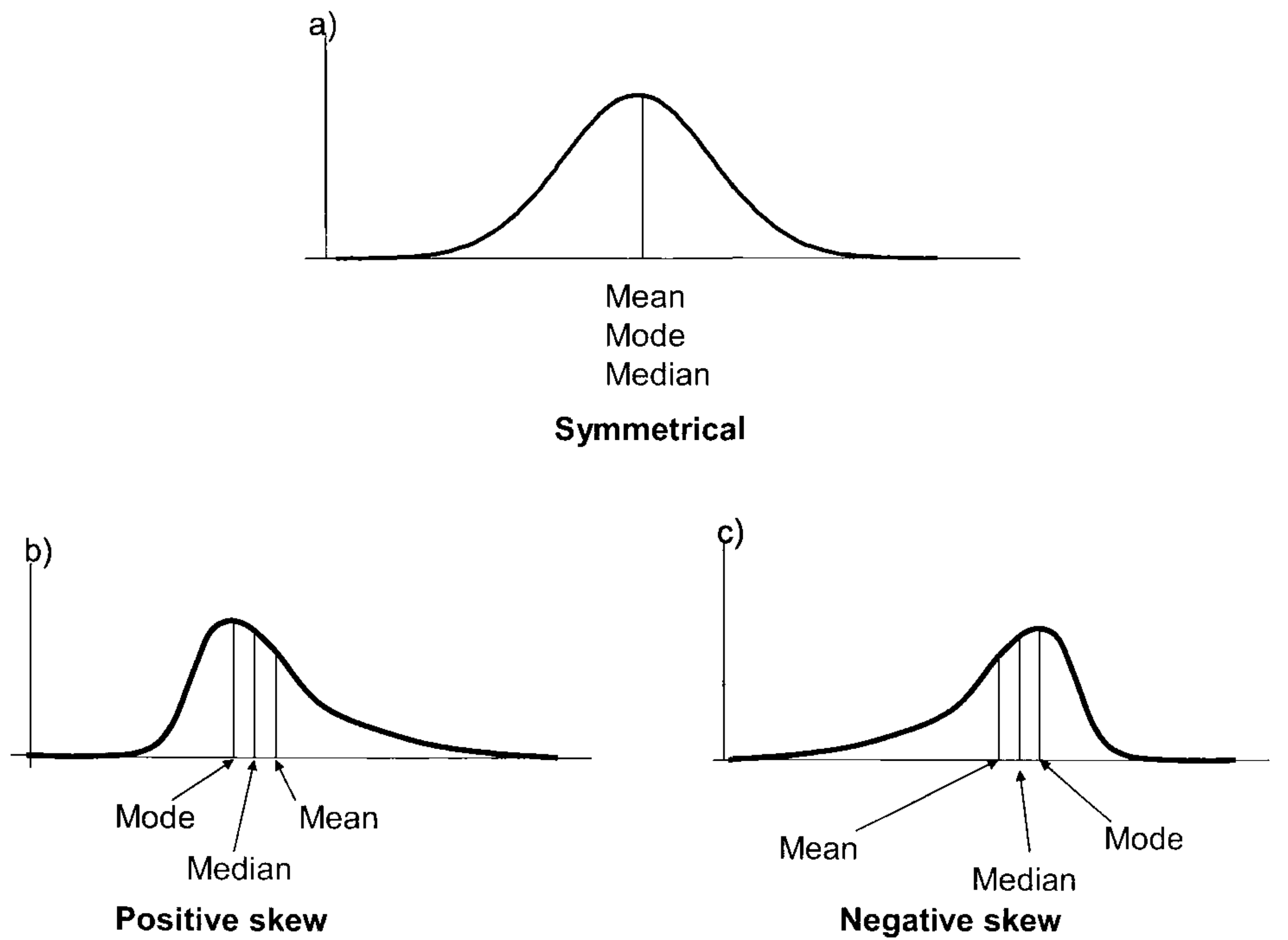


Verteilungsformen (nach Bortz 2006)

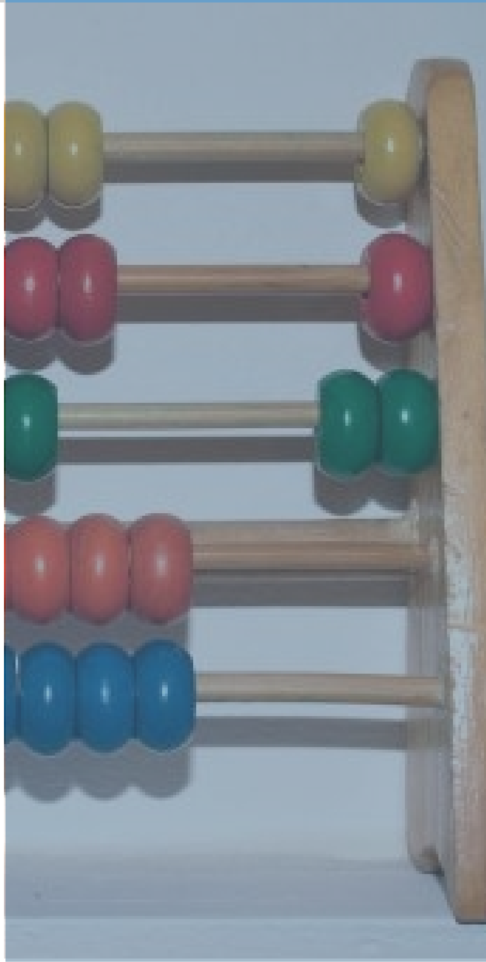
rechtsschief



linksschief


a)
Mean
Mode
Median
Symmetrical
b)
Mode
Median
Mean
Positive skew
c)
Mean
Median
Mode
Negative skew

Form der Verteilung [2]



Schiefe

Mittelwert rechts oder links vom Median

Ablesen aus dem Diagramm ;-)

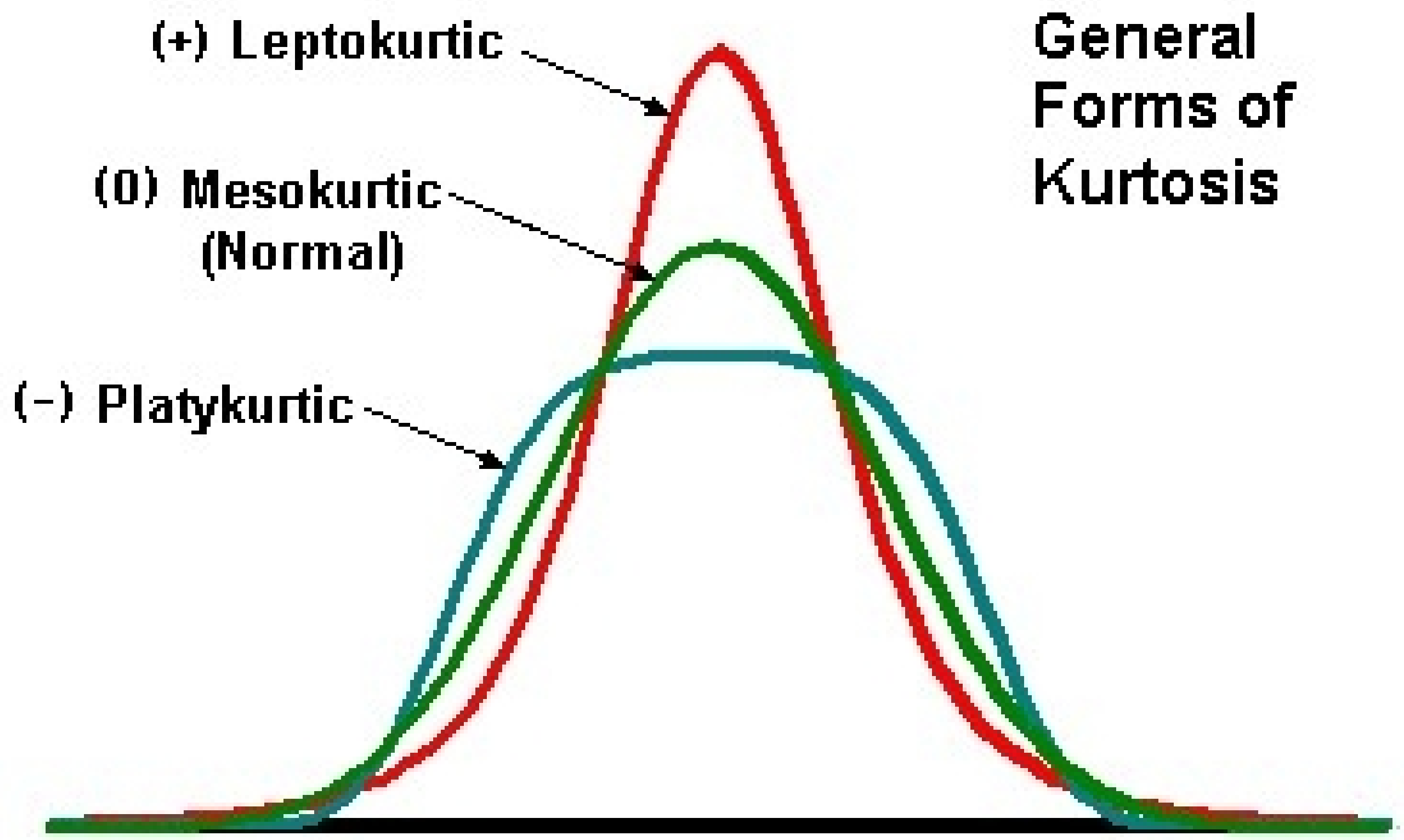
Berechnen:

$$\hat{S}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})^3}{n*s^3}$$

Positive bei Linksteil(rechtsschief), negativ bei rechtsteil(linksschief)

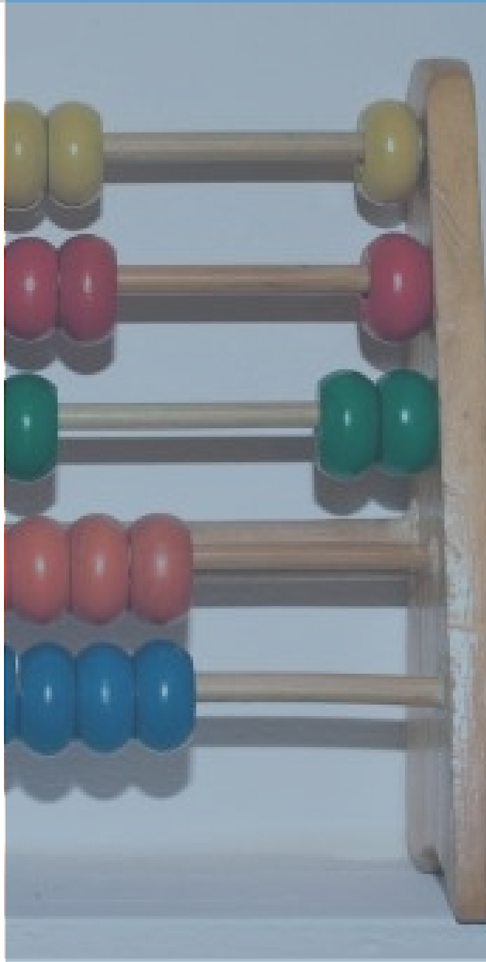
In R:

```
schiefe <- function (x) {
m3 <- sum((x-mean(x))^3) #Zähler
skew <- m3 / ((sd(x)^3)*length(x)) #Nenner
skew}
> test<-c(1,1,1,1,1,1,1,1,1,1,2,3,4,5)
> schiefe(test)
[1] 1.406826
> test<-c(3,3,3,3,3,3,3,3,3,3,3,3,2,1)
> schiefe(test)
[1] -2.231232
```



Quelle: http://grants.hhp.coe.uh.edu/doconnor/PEP6305/Topic%20002%20Organizing%20Data2.3.htm

Form der Verteilung [3]



Kurtosis

Die Wölbung der Verteilung

Ablesen aus dem Diagramm ;-)

Berechnen:

$$K=\frac{\sum_{i=1}^{n}(x_i-\bar{x})^4}{n*s^4}-3$$

Positive bei steilerem, negativ bei flacherem Anstieg als bei der Normalverteilung

In R:

```
> kurtosis <- function (x) {
m3 <- sum((x-mean(x))^4)
skew <- m3 / ((sd(x)^4)*length(x))-3
skew}
> test<-c(1,2,3,4,4,5,6,7)
> kurtosis(test)
[1] -1.46875
> test<-c(1,2,3,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,5,6,7)
> kurtosis(test)
[1] 2.011364
```

Nächste Sitzung 25. November 2010:
Nicht-parametrische Tests

Bitte lesen Sie Shennan Kapitel 4