C A U
Christian-Albrechts-Universität zu Kiel

03_explorative_statistik-graphische_darstellung

Tabellen und Diagramme

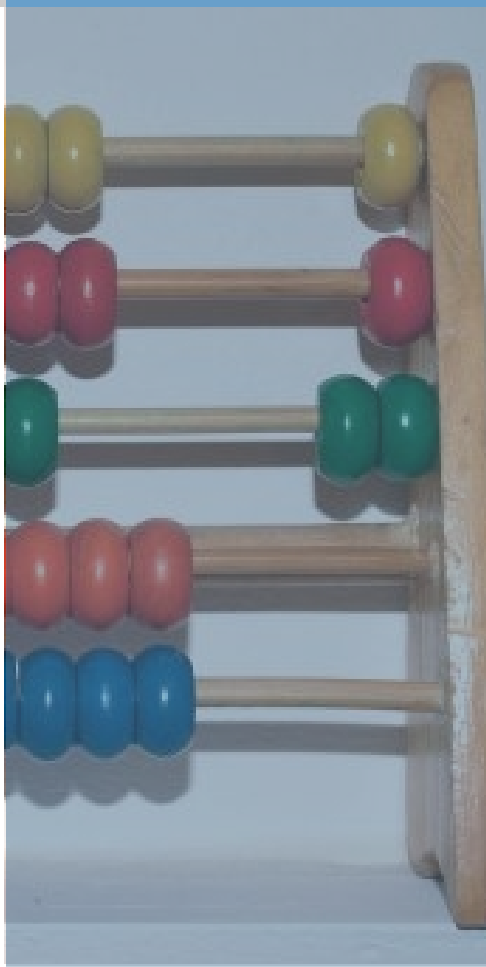
Grundlegende statistische Verfahren für archäologische Datenanalyse in R

Christian-Albrechts-Universität zu Kiel

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
```

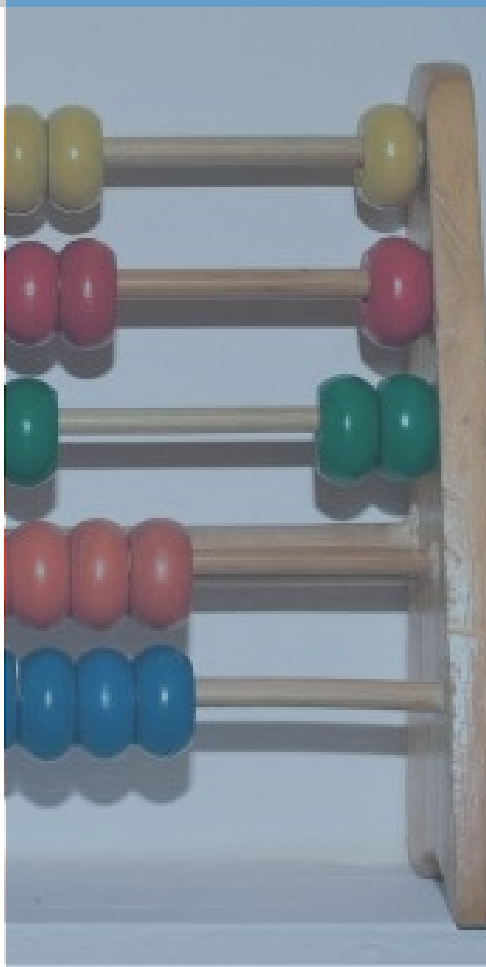


C | A | U
Christian-Albrechts-Universität zu Kiel

Kreuztabellen (Kontingenztafel)

Dient zur Zusammenfassung von z.B. Daten:

```
> tabelle<-table(laender$einlieferer,laender$Kontinent)
> tabelle

           Afrika Asien Europa Nordamerika
  breske        0     0      2            0
  eberle        1     1      1            0
  frank         0     1      2            0
  greve         0     0      3            0
  lublasser     0     3      0            0
  wiese         0     0      0            1
> addmargins(tabelle)

           Afrika Asien Europa Nordamerika Sum
  breske        0     0      2            0   2
  eberle        1     1      1            0   3
  frank         0     1      2            0   3
  greve         0     0      3            0   3
  lublasser     0     3      0            0   3
  wiese         0     0      0            1   1
  Sum           1     5      8            1  15
```



Dient zur Zusammenfassung von z.B. Daten:

```
> ftable(laender$einlieferer~laender$kontinent)
                  laender$einlieferer breske eberle frank greve lublasser wiese
laender$kontinent
Afrika                                     0      1     0     0          0     0
Asien                                      0      1     1     0          3     0
Europa                                     2      1     2     3          0     0
Nordamerika                                0      0     0     0          0     1
```

C A U
Christian-Albrechts-Universität zu Kiel

Grundsätzliches zu Diagrammen

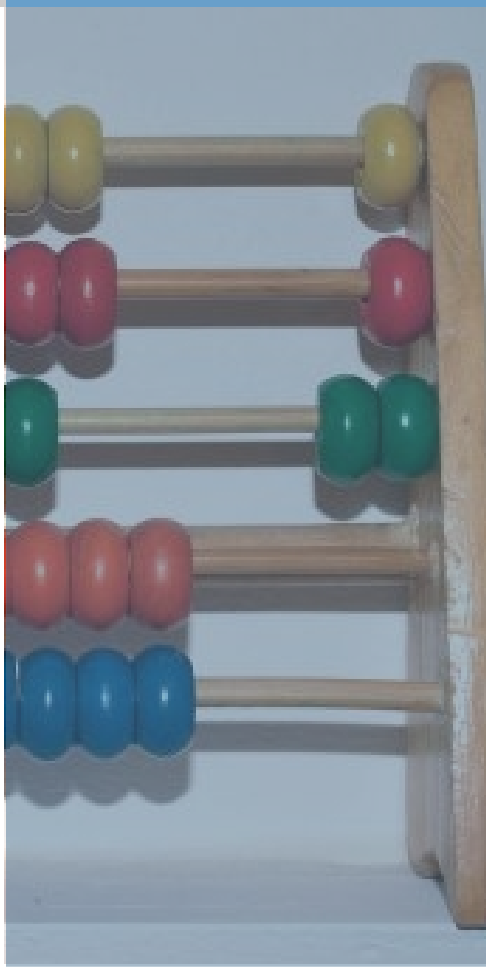
Grundsätze guter Diagramme nach E. Tufte:
(The Visual Display of Quantitative Information. Cheshire/ Connecticut: Graphics Press, 1983)

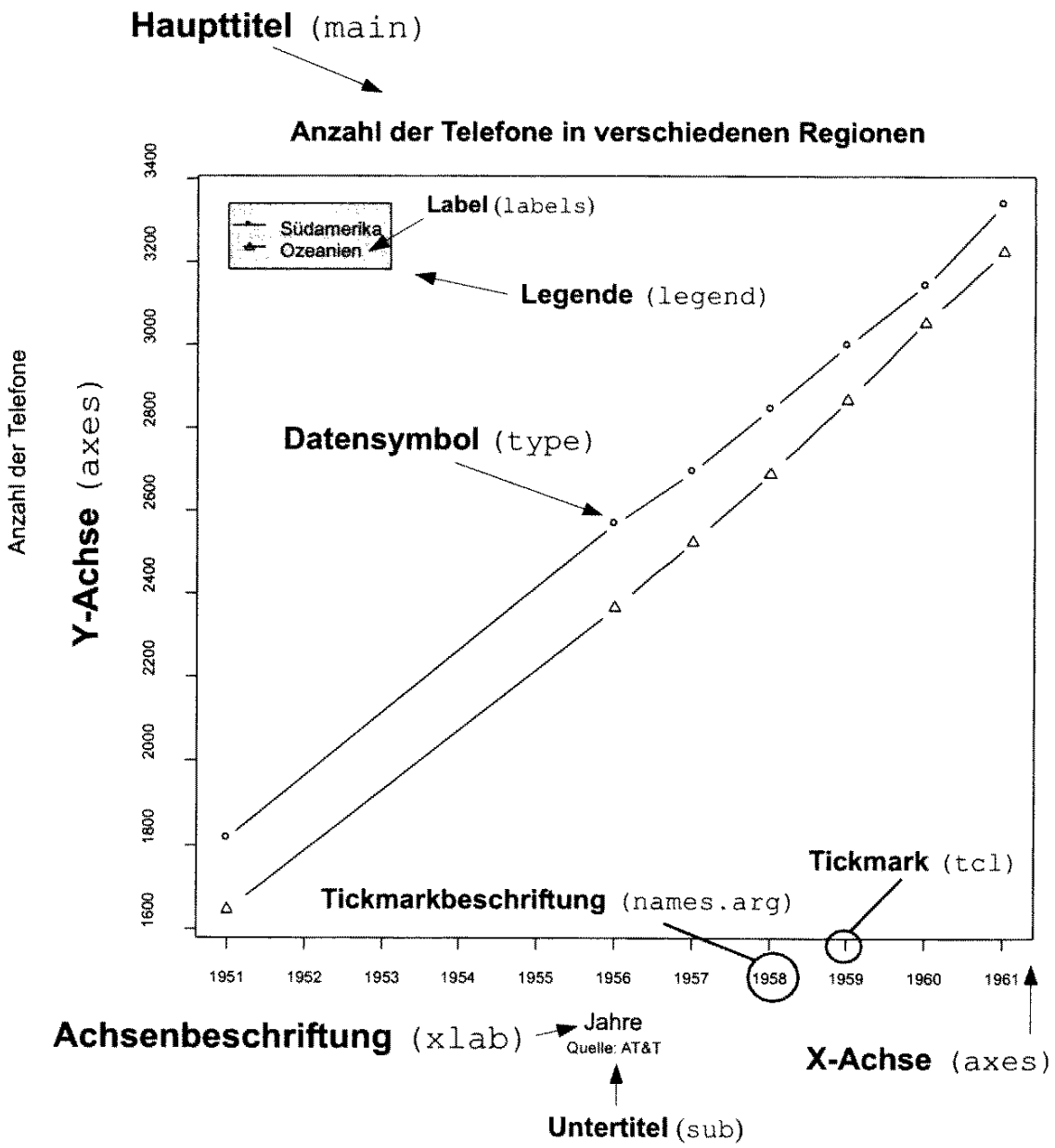
• „Graphical exellence is that which gives to the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space."

• Data-ink ratio = „proportion of a graphic's ink devoted to the non-redundant display of data-information" (kein chartjunk!)

• „Graphical excellence is often found in simplicity of design and complexity of data."

Nach Müller-Scheeßel




Haupttitel (main)
Anzahl der Telefone in verschiedenen Regionen
Label (labels)
Südamerika
Ozeanien
Legende (legend)
Datensymbol (type)
Anzahl der Telefone
Y-Achse (axes)
3400
3200
3000
2800
2600
2400
2200
2000
1800
1600
Tickmark (tcl)
Tickmarkbeschriftung (names.arg)
1951
1952
1953
1954
1955
1956
1957
1958
1959
1960
1961
Achsenbeschriftung (xlab)
Jahre
Quelle: AT&T
X-Achse (axes)
Untertitel (sub)

C | A | U
Christian-Albrechts-Universität zu Kiel

Plot [1]

Grundlegende Zeichenfunktion von R:

```
> plot(laender$Einwohnerzahl)
```

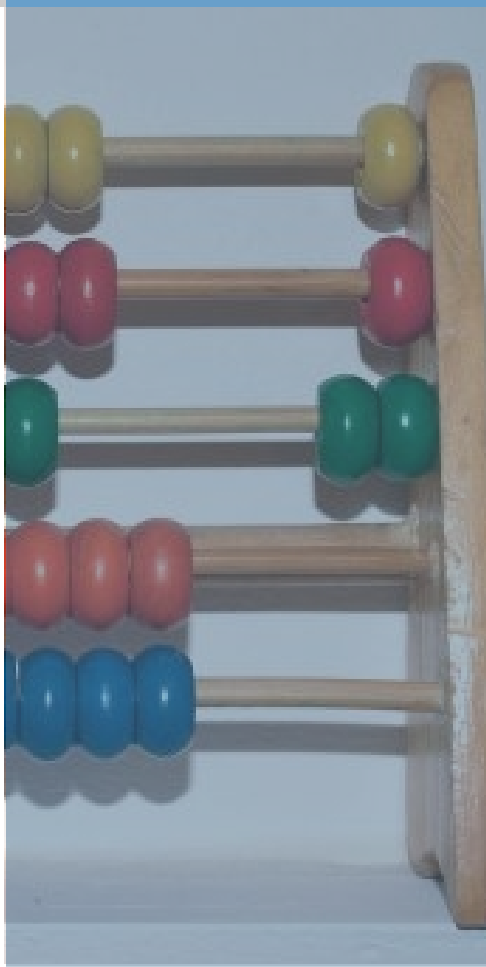
Weitere Optionen:

- p – Punkte (Voreinstellung)
- l – durchgezogene Linie
- b – Linie mit Punkten bei den Werten
- c – Linie mit Lücken bei den Werten
- o – durchgezogene Linie mit Punkten bei den Werten
- h – vertikale Linie bis zu den Werten
- s – gestufte Linie von Wert zu Wert
- n – leeres Koordinatensystem

```
> plot(laender$Einwohnerzahl,type="b")
```

Intelligentes System: automatische Bestimmung des Variablentyps, Anlage eines passenden Diagramms

```
> plot(laender$kontinent)
```



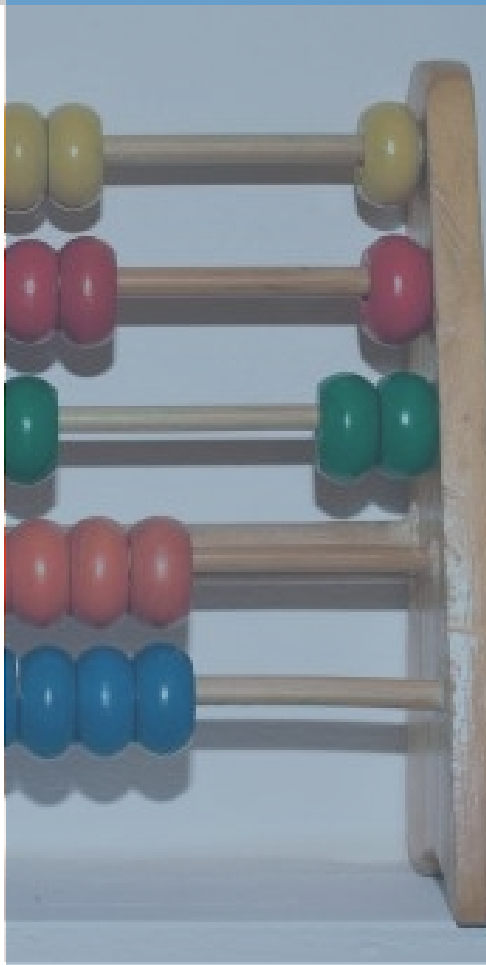
Grundlegende statistische Verfahren für archäologische Datenanalyse in R

Christian-Albrechts-Universität zu Kiel

Plot [2]

Optionen und Beschriftungen einfügen:

```
> plot(laender$Fläche,laender$Weltrang.CPI,
     xlim=c(0,2500000), # Grenze der X-Achse
     ylim = c(0,200), # Grenze der Y-Achse
     ylab = "Weltrang nach CPI", # Beschriftung der y-achse
     xlab = "Fläche", # Beschriftung der x-achse
     main = "Fläche vs. Weltrang nach BPI", # Titel des Diagramms
     sub="Beispielgraphik" #Untertitel des Diagramms
)
```



C A U
Christian-Albrechts-Universität zu Kiel

Plot [3]

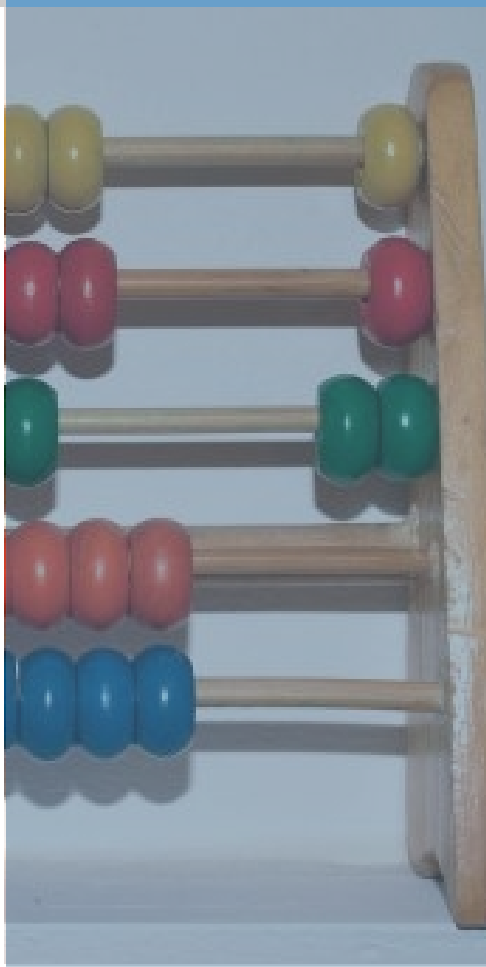
Plot tut einiges für Sie:

Öffnen eines Fensters für die Ausgabe
Bestimmen der optimalen Größe für den Bezugsrahmen
Einzeichnen des Koordinatensystems
Zeichnen der Werte

Erstellen eines „handles“ für weitere Ergänzungen des Plots durch z.B.:

lines	– zeichnet Linien in eine vorhandene Graphik
points	– zeichnet Punkte in eine vorhandene Graphik
abline	– zeichnet spezielle Linien in eine vorhandene Graphik
text	– zeichnet Text an beliebiger Stelle in eine vorhandene Graphik

Weitere Möglichkeiten der Gestaltung: ? par



Christian-Albrechts-Universität zu Kiel

Plot [4]

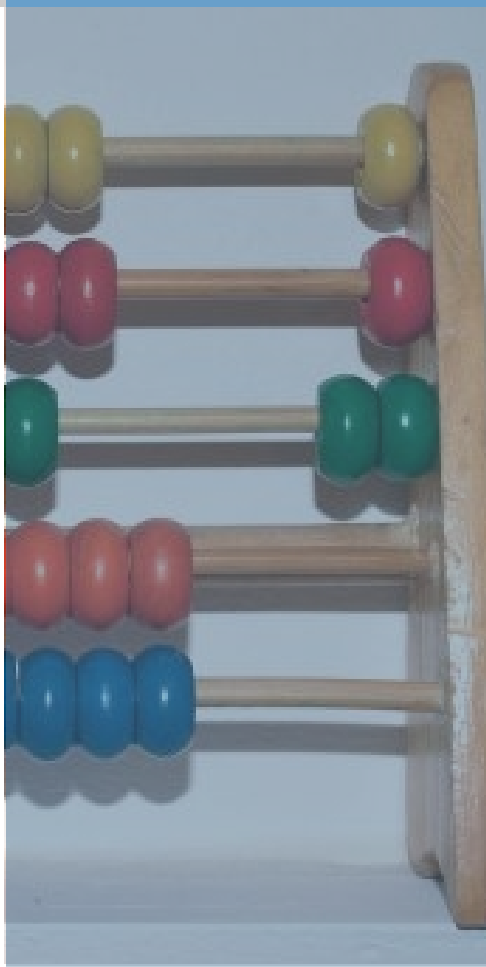
Zusätzliche Elemente einfügen:

Linien einzeichnen

```
> abline(v=mean(laender$Fläch,na.rm=T))
> abline(h=mean(laender$Weltrang.CPI,na.rm=T))
>abline(lm(laender$Weltrang.CPI~laender$Fläche.in.km))
```

Text einzeichnen

```
> text(2000000, mean(laender$Weltrang.CPI), # Position bei x 20 und y arithm Mittel erhalten
label = paste("MW (Ösen erhalten)= ", # Text besteht
round(mean(daten$Oese.zahl.erhalten,na.rm=T))),
# aus Zusammensetzung, die mittels paste() verbunden wird
pos = 3, # Positionierung oberhalb
cex = 0.7 # Schriftgröße 70%
)
```



Christian-Albrechts-Universität zu Kiel

Speichern von Graphiken

Mittels der GUI:
File → Save as...

Mittels der Kommandozeile:
Als Vektor-Graphik

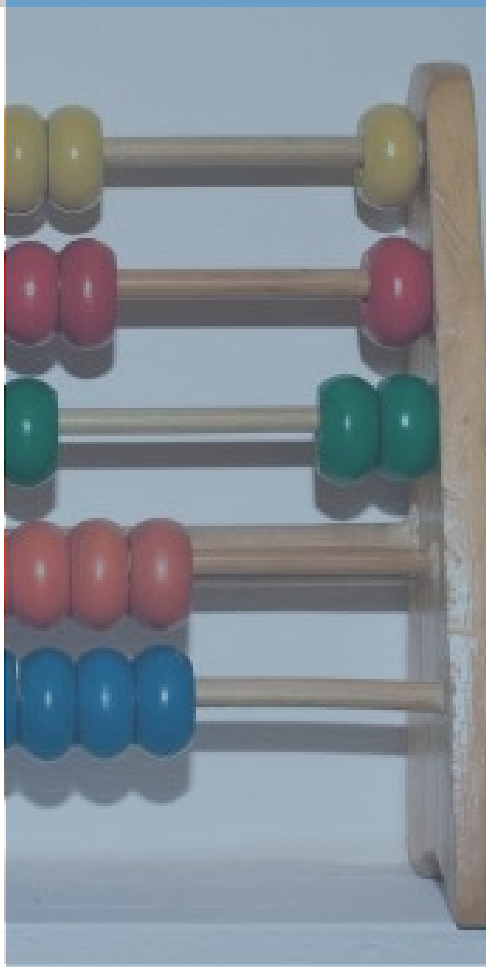
```
> dev.copy2eps(file="test.eps")
> dev.copy2pdf(file="test.pdf")
```

Als Bitmap-Datei

```
> savePlot(filename="test.tif", type="tiff")
```

Möglich sind “png”, “jpeg”, “tiff”, “bmp”

SavePlot beherrscht teilweise auch Vektor-Graphiken (hängt vom Betriebssystem und der Installation ab...



C | A | U
Christian-Albrechts-Universität zu Kiel

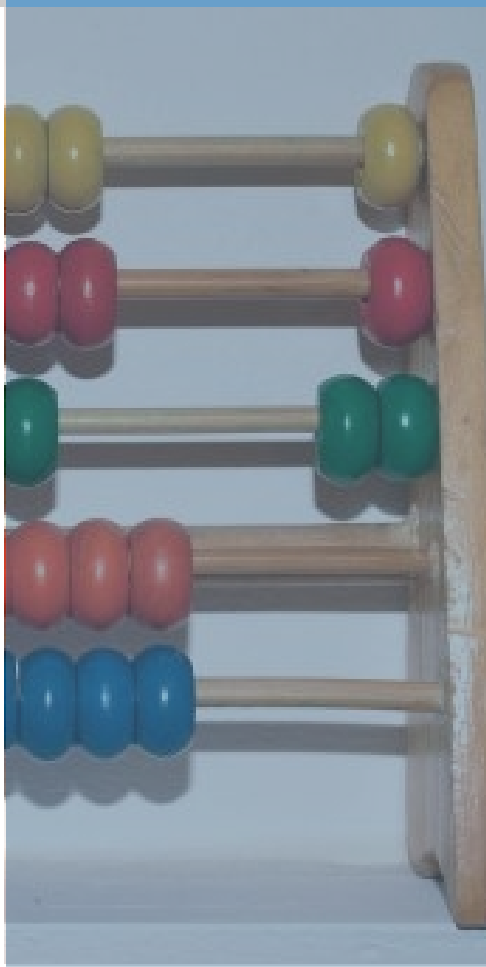
Tortendiagramm (Kreisdiagramm) [1]

Der Klassiker – mit R auch nicht viel besser...
Dient zum Darstellen von Anteilen, geeignet für nominale Daten

$$a_i = \frac{h_i}{N} \cdot 360°$$

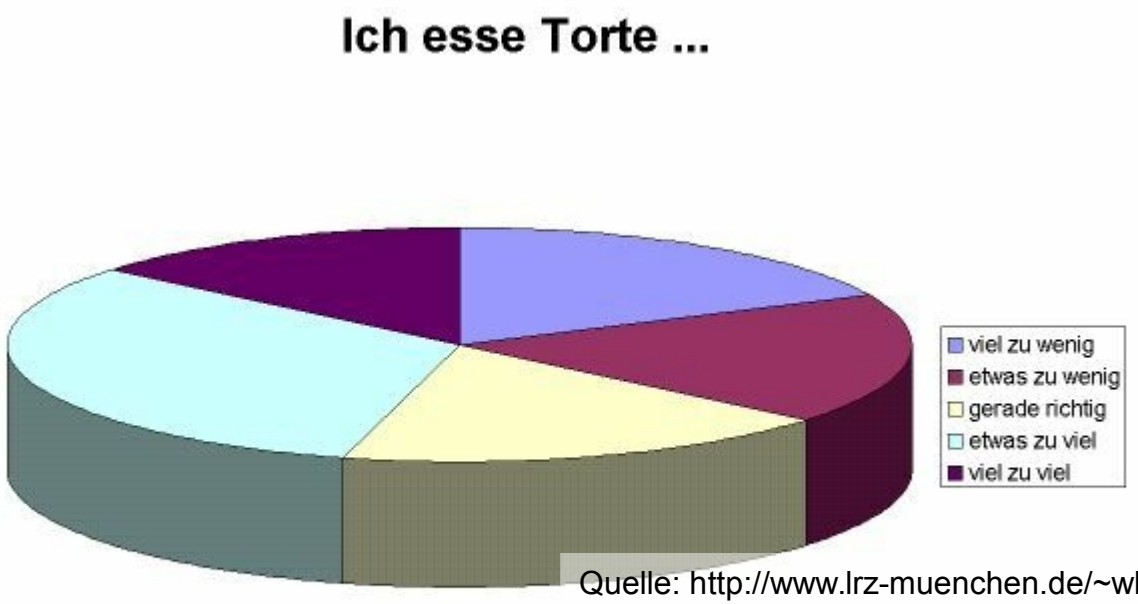
Nachteile:
Farbwahl kann Wahrnehmung beeinflussen (rot wird größer gesehen als grau)
Kleinere Unterschiede sind schlecht wahrnehmbar

Völliges No-Go: dreidimensionale Torten!!!



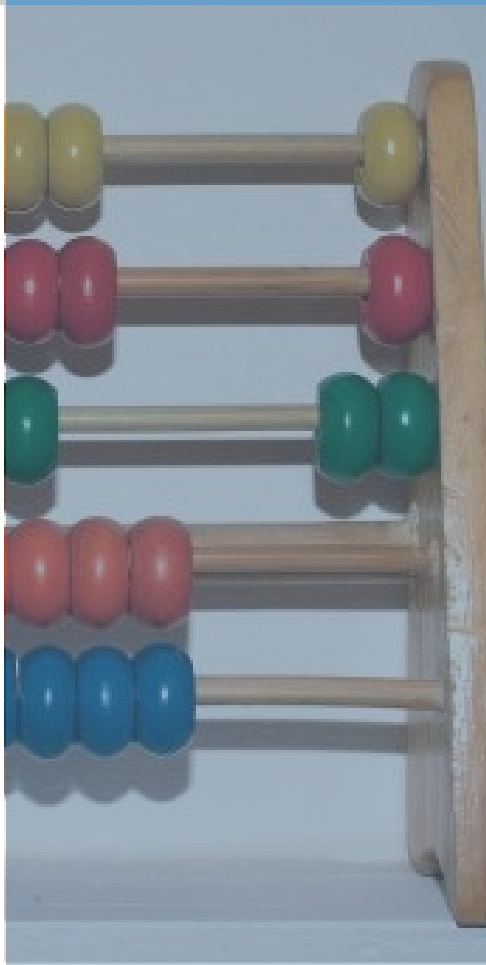
C | A | U
Christian-Albrechts-Universität zu Kiel

Tortendiagramm (Kreisdiagramm) [2]



Quelle: http://www.lrz-muenchen.de/~wlm

Die Stücke »viel zu wenig«, »etwas zu wenig« und »gerade richtig« sind genau gleich groß, das Stück »viel zu viel« ist noch etwas kleiner.



C A U
Christian-Albrechts-Universität zu Kiel

Tortendiagramm (Kreisdiagramm) [3]

Torten in R

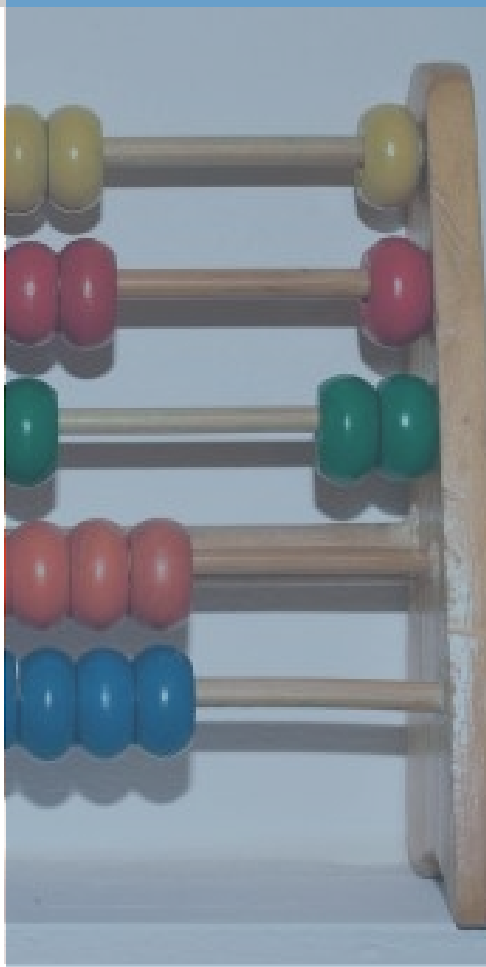
Eingabe ist ein Vektor von Anzahlen

```
> table(laender$kontinent)

     Afrika       Asien      Europa Nordamerika
          1           5           8           1
> pie(table(laender$kontinent))
```

Farbpaletten:

Die Standard-Farbpalette ist pastell, wenn man eine andere möchte:

```
> pie(table(laender$kontinent), col=c("red","green","blue","yellow"))
```



C | A | U
Christian-Albrechts-Universität zu Kiel

Säulendiagramm [1]

Meist die bessere Alternative

Säulendiagramme eignen sich zur Darstellung von Proportionen wie auch für absolute Daten. Sie können für alle Skalenniveaus eingesetzt werden.

```
> barplot(table(laender$kontinent))
> windows() # öffnet neues Fenster, unter linux x11(), unter mac
quartz ()
> barplot(laender$Fläche.in.km.)
```

Mit Namen:

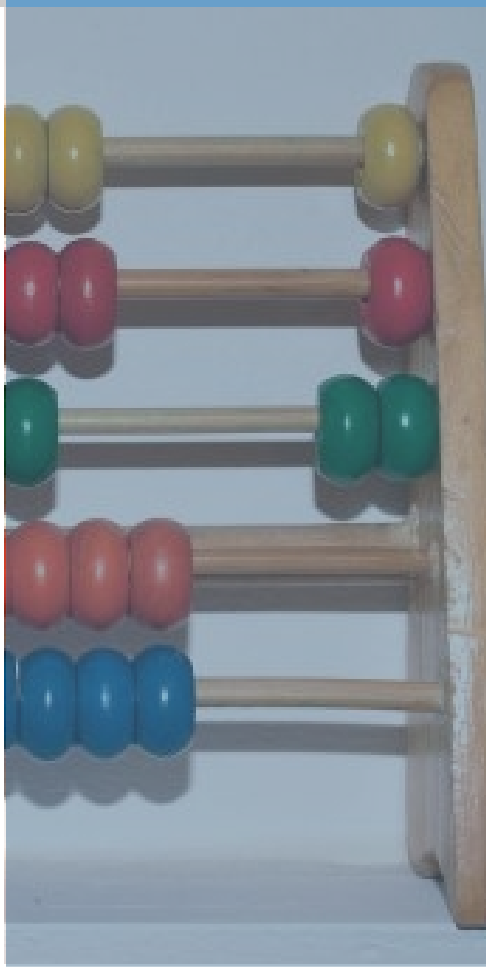
```
> par(las=2)
> barplot(laender$Fläche.in.km.,names.arg=laender$Name)
```

Mit Überschrift:

```
> title("Fläche der Sample-Länder")
> par(las=1)
```

Horizontal:

```
> barplot(table(laender$kontinent),horiz=T,cex.names=0.5)
```

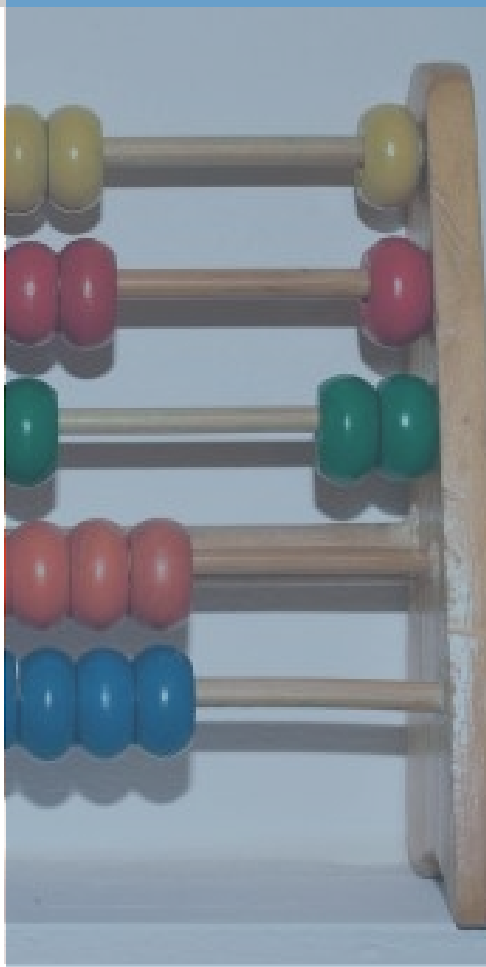


Christian-Albrechts-Universität zu Kiel

Säulendiagramm [2]

Darstellung von Anteilen absolut

```
> tabelle
          Afrika Asien Europa Nordamerika
  breske       0     0      2           0
  eberle       1     1      1           0
  frank        0     1      2           0
  greve        0     0      3           0
  lublasser    0     3      0           0
  wiese        0     0      0           1

> barplot(tabelle)

> barplot(tabelle, beside=T)

> barplot(tabelle, beside=T, legend.text=T)

> barplot(tabelle, beside=T, legend.text=T, ylim=c(0,5))

> barplot(tabelle, beside=T, legend.text=T, xlim=c(0,36))
```

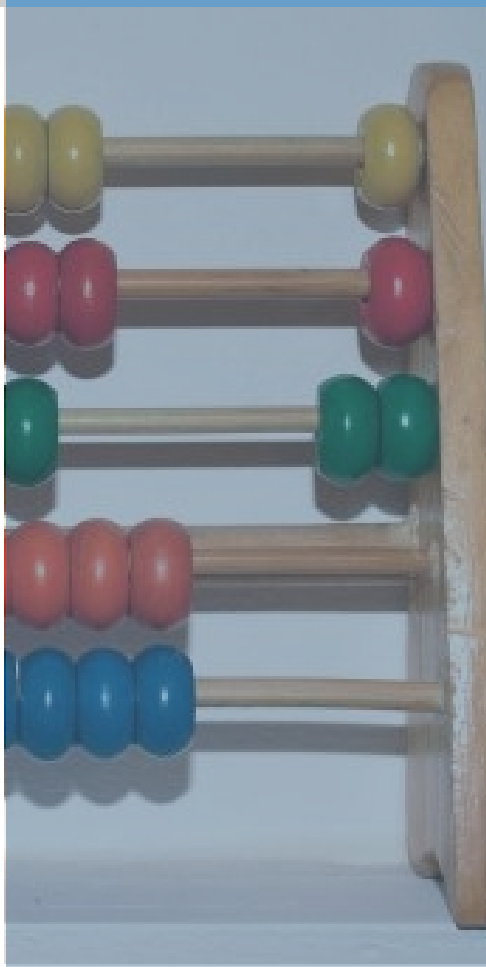


Christian-Albrechts-Universität zu Kiel

Säulendiagramm [3]

Darstellung von Anteilen relativ

```
> tabelle.prop<-prop.table(tabelle,2)
> tabelle.prop

           Afrika Asien Europa Nordamerika
  breske    0.000 0.000  0.250       0.000
  eberle    1.000 0.200  0.125       0.000
  frank     0.000 0.200  0.250       0.000
  greve     0.000 0.000  0.375       0.000
  lublasser 0.000 0.600  0.000       0.000
  wiese     0.000 0.000  0.000       1.000
> barplot(tabelle.prop)

> tmp<-barplot(tabelle.prop, legend.text=T, col=rainbow(11),
xlim=c(0,8))

> title("Anteile der Einlieferer \n je Kontinent", outer=TRUE,
line=-3)
```



Christian-Albrechts-Universität zu Kiel

Säulendiagramm [4]

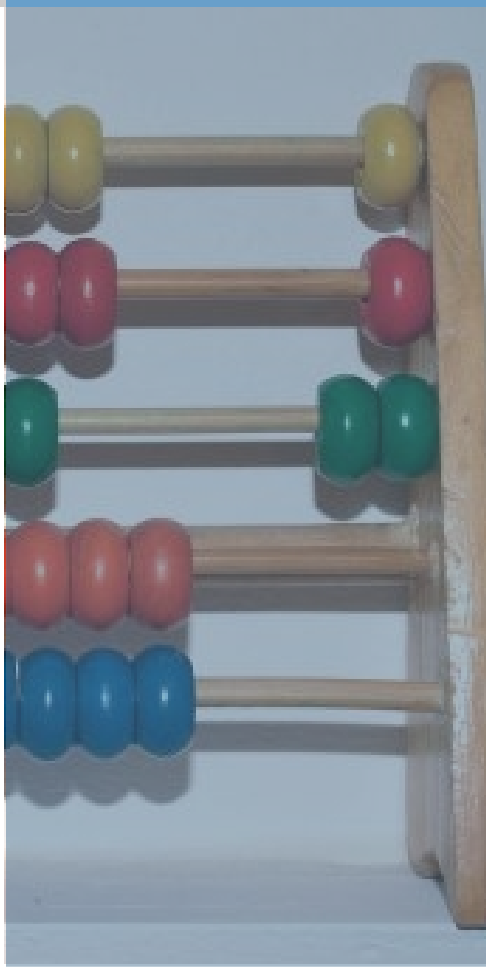
Probleme mit Säulen – und auch vielen anderen – Diagrammen

Prozente vs. Anzahl: Prozente verzerren häufig die Verhältnisse

```
> par(mfrow=c(2,1))

> barplot(tabelle,beside=T)
> barplot(tabelle.prop,beside=T)
```

Skalen: die gewählte Reichweite der Achsen kann die Verhältnisse verzerren

```
> par(mfrow=c(1,2))

> barplot(laender$Fläche.in.km.[c(2,3)],xpd=F,ylim=c(250000,500000))
> barplot(laender$Fläche.in.km.[c(2,3)],xpd=F)

>par(mfrow=c(1,1))
```



Christian-Albrechts-Universität zu Kiel

Box-plot (Box-and-Whiskers-Plot)

Einer der besten!

Dient zur Darstellung der Verteilung von Werten innerhalb einer Datenreihe von metrischen Daten (intervall, verhältnis)

```
1  2  3  4  5  6  7  8  9
____|___|___|____
```

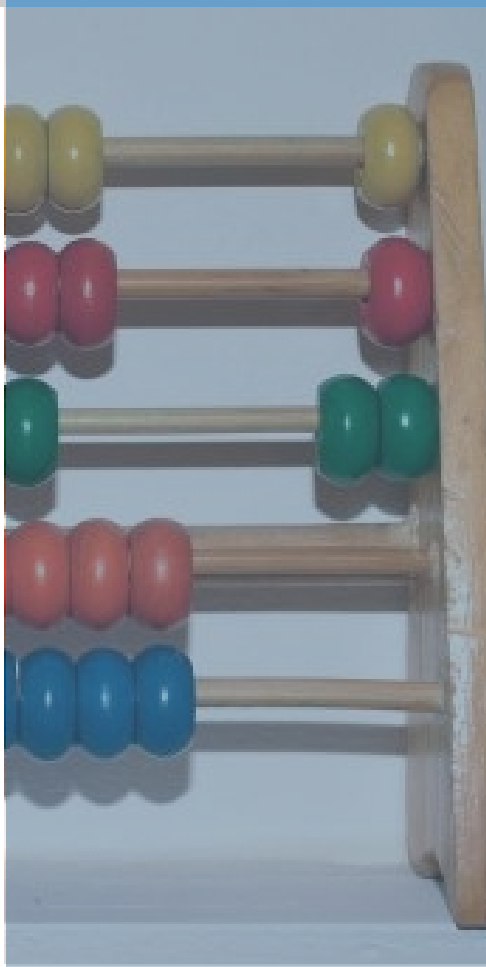
```
> boxplot(1:9)
```

Box: die inneren beiden Quantile
Whisker: letzter Wert < als 1.5-fache des Abstands vom Quantil zum Median

```
> boxplot(laender$Fläche)
> boxplot(laender$Fläche.in.km.~laender$einlieferer)
```

In hübsch:

```
> par(las=1)
> boxplot(laender$Fläche.in.km.~laender$einlieferer, data = daten,
  main = "Fläche der Länder \n nach Einlieferer", col="grey",
  xlab="Einlieferer", ylab= "Fläche")
```



Christian-Albrechts-Universität zu Kiel

Streudiagramm (scatterplot)

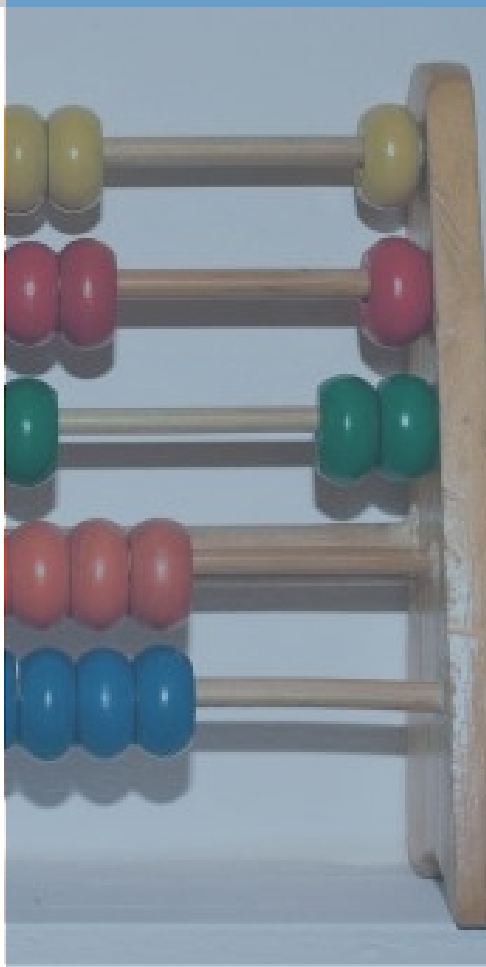
Für 2 diskrete Variablen

Dient zur Darstellung einer Variable in Abhängigkeit zu einer anderen. Prinzipiell sind alle Skalierungen möglich, für nominale und ordinale gibt es aber meist bessere.

```
> plot(laender$Weltrang.CPI,laender$Fläche.in.km.)
>
abline(lm(laender$Fläche.in.km.~laender$Weltrang.CPI),
col="red")
```

Weitere Bibliotheken aufrufen:

```
> library(car) # bibliothek zur Regressionsanalyse
> scatterplot(Fläche.in.km.~Weltrang.CPI,data=laender)

> library(ggplot2)
> b<-
ggplot(laender,aes(x=Weltrang.CPI,y=Fläche.in.km.))
> graph<-b + geom_point()
> show(graph)
```



C A U
Christian-Albrechts-Universität zu Kiel

Liniendiagramm

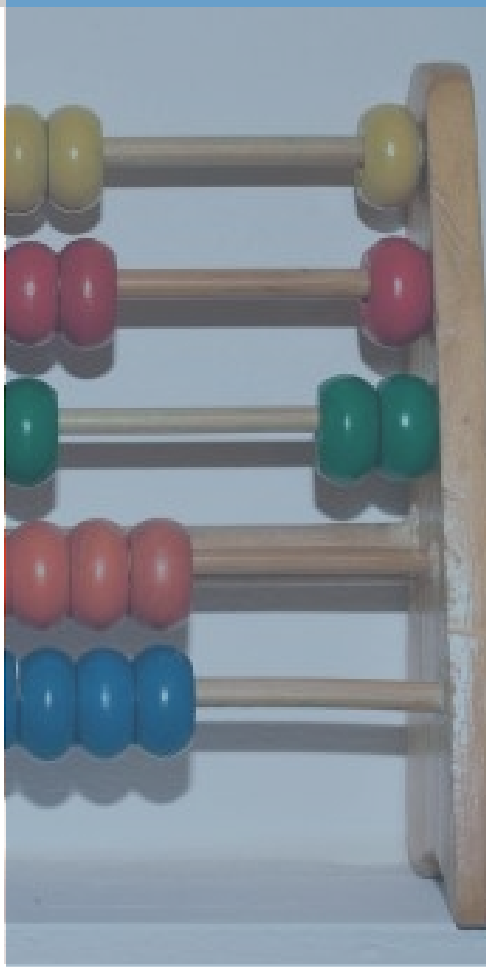
Für 2 stetige Variablen bei kontinuierlichen Vorgängen
Dient zur Darstellung einer Variable in Abhängigkeit zu einer anderen.
Gleiches gilt wie für den scatterplot.

Nicht sehr sinnvoll:

```
> plot(laender$Weltrang.CPI,laender$Fläche.in.km., type="n")
> lines(sort(Fläche.in.km.)~sort(Weltrang.CPI), data = laender)
```

sinnvoller:
http://www.radiocarbon.org/IntCal04%20files/IntCal04_rawdata.csv
Speichern im Arbeitsverzeichnis

```
> intcal<-read.csv("IntCal04_rawdata.csv")

> plot(intcal$Starting.cal.age.BP,intcal$X14C.age.BP, type="l", xlim=c(0,100), ylim=c(0,200))
> lines(intcal$Starting.cal.age.BP, intcal$X14C.age.BP+intcal$total.14C.uncertainty, col="lightgrey", lty=2)
> lines(intcal$Starting.cal.age.BP, intcal$X14C.age.BP-intcal$total.14C.uncertainty, col="lightgrey", lty=2)
```



Christian-Albrechts-Universität zu Kiel

Histogramm

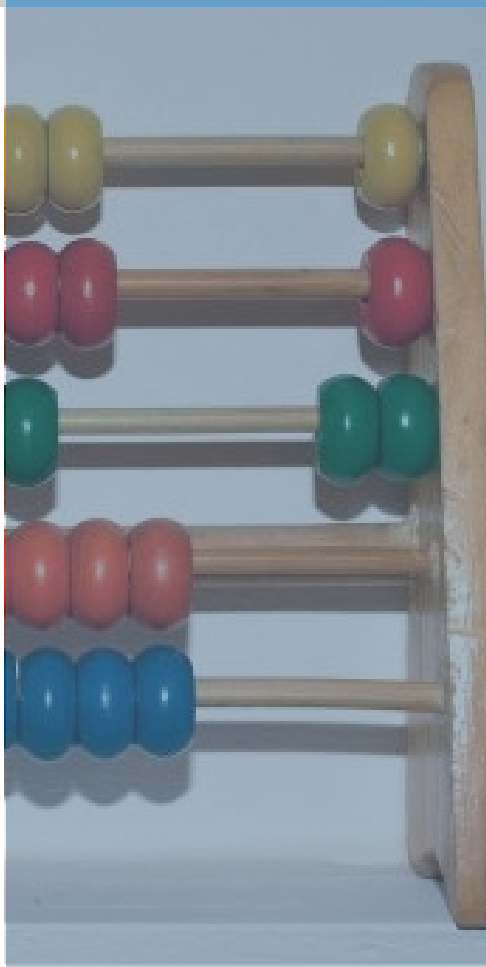
Dient zur klassifizierten Darstellung von Verteilungen
Datenreduktion vs. Genauigkeit. Darstellung von aus metrischen (stetigen) Variablen gewonnenen Zählwerten (Absolutskala).

```
> hist(laender$Fläche)
> hist(laender$Fläche, labels=T)
> hist(laender$Fläche, labels=T, breaks=20)
```

In hübsch

```
> hist(laender$Fläche,breaks=20,labels=T, col="red",
xlab="Fläche", main="Histogramm der Fläche
ausgewählter Länder")
```

Nachteile:
Datenreduktion vs. Genauigkeit → Informationsverlust
Darstellung hängt sehr stark von der Wahl der Klassenbreite ab



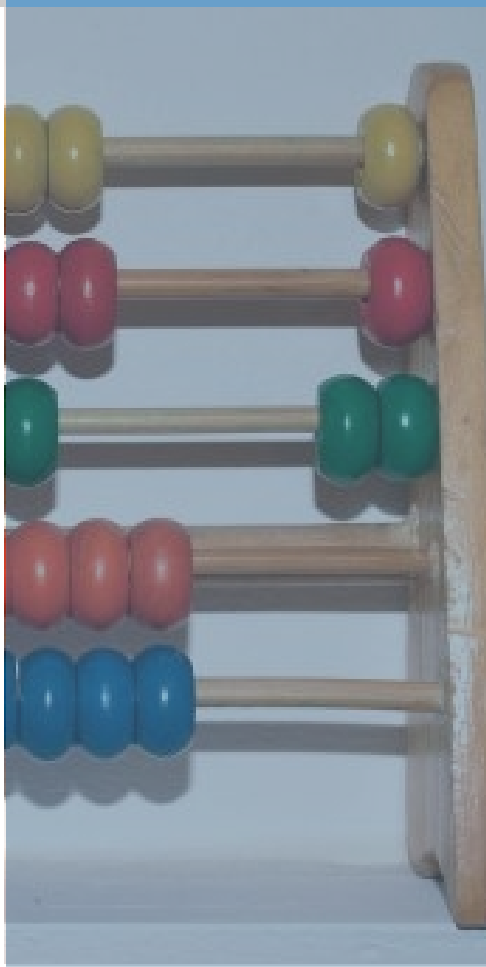
Christian-Albrechts-Universität zu Kiel

steam-and-leaf Diagramm

Versucht, Nachteile eines Histogramms zu vermeiden
Hat sich aber kaum durchgesetzt. Skala wie Histogramm.

```
> stem(laender$Fläche.in.km.)

  The decimal point is 6 digit(s) to the right of the |

  0 | 00000001344467
  2 | 2
  4 |
  6 |
  8 | 8
```

Vorteil: Die Informationen über die Verteilung innerhalb der Klassen und die absoluten Werte werden mit angezeigt.



Christian-Albrechts-Universität zu Kiel

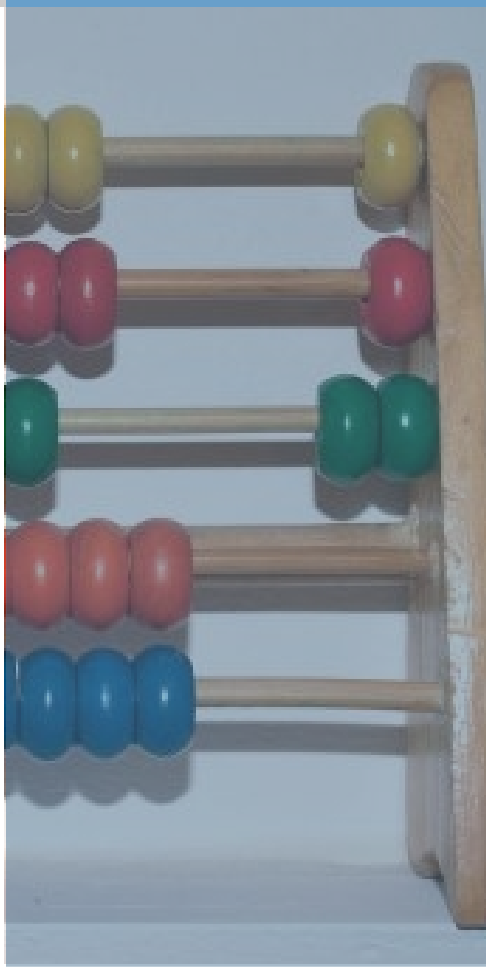
kernel smoothing (kernel density estimation)

Ein anderer Versuch, die Nachteile von Histogrammen zu umgehen
Die Verteilung der Werte wird berücksichtig und eine Verteilungskurve wird berechnet. Gleichmäßige Verteilungen werden besser, ohne künstliche Brüche, wiedergegeben. Skala wie Histogramm

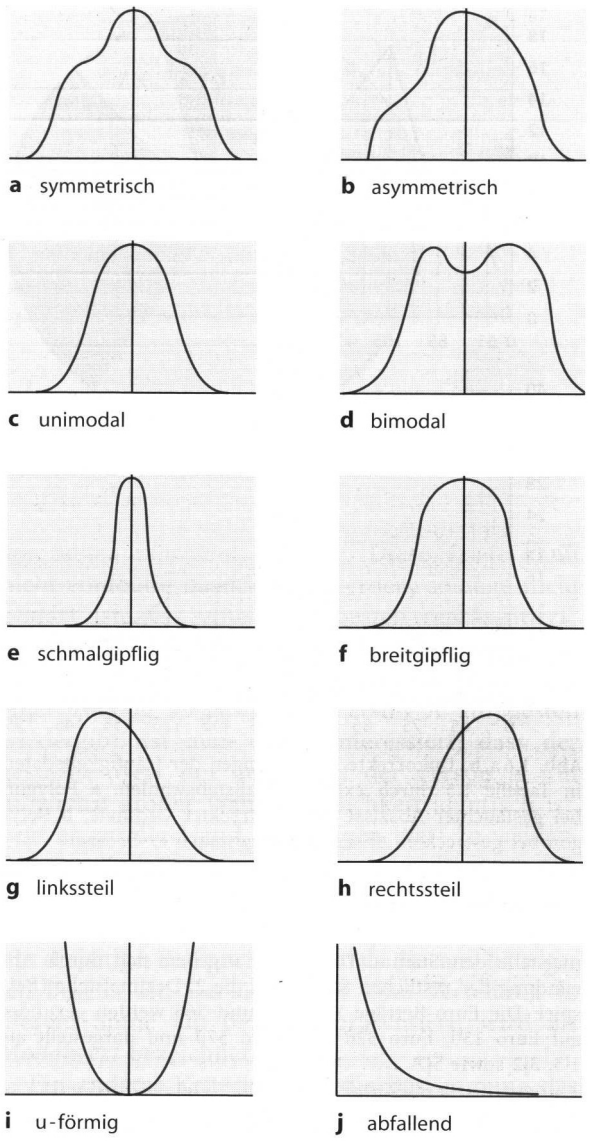
```
> plot(density(laender$Fläche))
```

Histogramm und kernel-density-plot auf einmal

```
> hist(laender$Fläche,breaks=20,labels=T, col="red",
xlab="Fläche", main="Histogramm der Fläche
ausgewählter Länder",prob=T)
> lines(density(laender$Fläche),lwd=4)
```



Verteilungsformen (nach Bortz 2006)



Christian-Albrechts-Universität zu Kiel

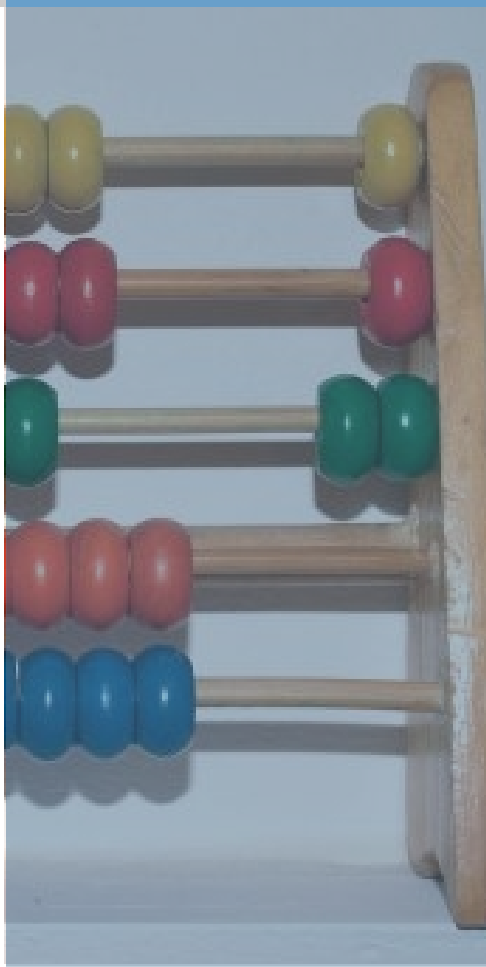
kulmulative Häufigkeitsverteilung

Darstellung der Anteile ordinal skalierter Variablen

Beispiel nach Shennan: Anzahl der Bestattungen nach Altersstufen

Infans I	Infans II	Juvenil	Adult	Matur	Senil
10	16	10	32	34	4

```
> bestattungszahl<-c(10,16,10,32,34,4)
> names(bestattungszahl)<-c("Infans I","Infans
II","Juvenil","Adult","Matur","Senil")
> plot(c(0, cumsum(bestattungszahl)/sum(bestattungszahl)), type="l",
axes="F", xlab="", ylab="Kulmulativer Anteil")
> axis(1,at=1:(length(bestattungszahl)+1),
c(0,names(bestattungszahl)))
> axis(2)
> box()
> title("Kulmulativer Anteil der Bestattungen nach Altersstufen")
```



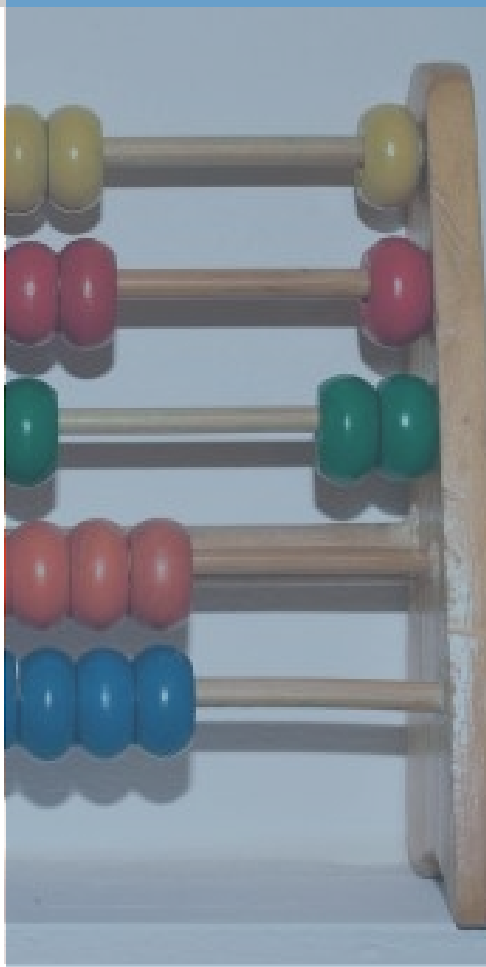
Christian-Albrechts-Universität zu Kiel

Dreiecksdiagramm

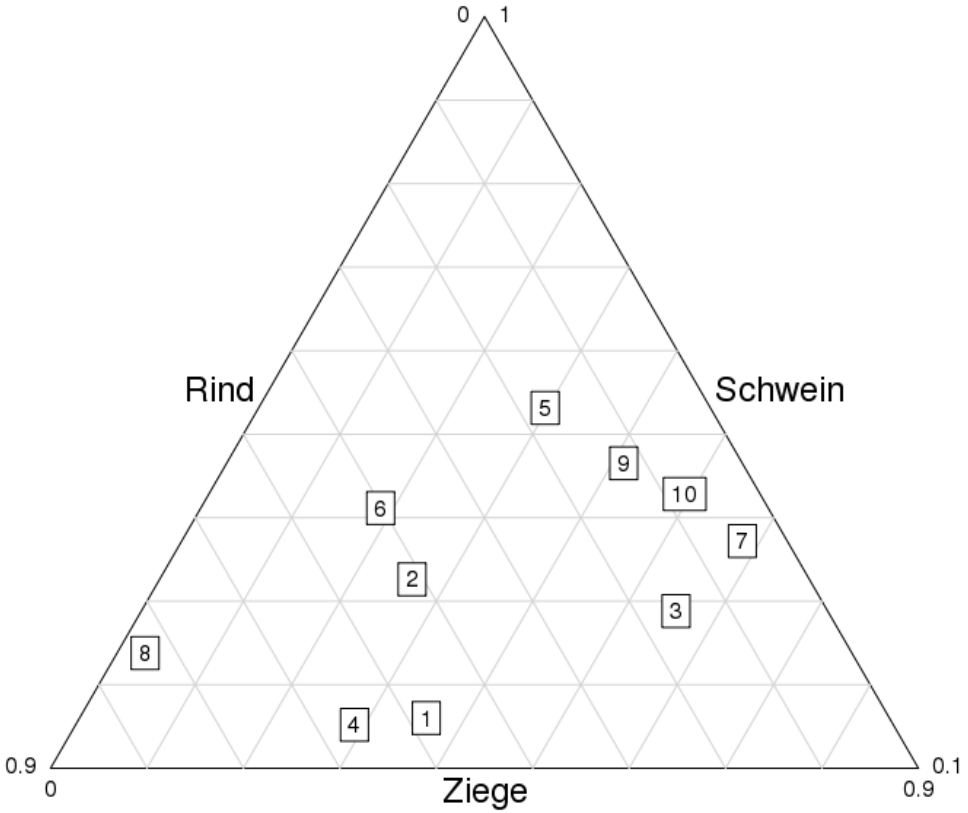
Einfachste Art eines multivariaten Diagramms

Dient zur Darstellung des Verhältnisses dreier Variablen zueinander. Prinzipiell für alle Skalierungen verwendbar, Daten werden in Prozentwerte umgerechnet...

```
> library(ade3)
> test<-matrix(round(abs(rnorm(30)*100)),ncol=3)
> colnames(test)<-c("Rind","Ziege","Schwein")
> test
      Rind Ziege Schwein
 [1,]  195   146      65
 [2,]   96    61      76
 [3,]   36   127      66
 [4,]  114    59      31
 [5,]   49    85     152
 [6,]  168    78     172
 [7,]   10   125      80
 [8,]  151     6      49
 [9,]   23    77      87
[10,]   48   303     263
> test<-as.data.frame(test)
> triangle.plot(test,label=rownames(test), clab =1, show=F,
labeltriangle=T)
```



Simuliertes Dreiecksdiagramm über die Verteilung von Tierknochen in verschiedenen Siedlungen

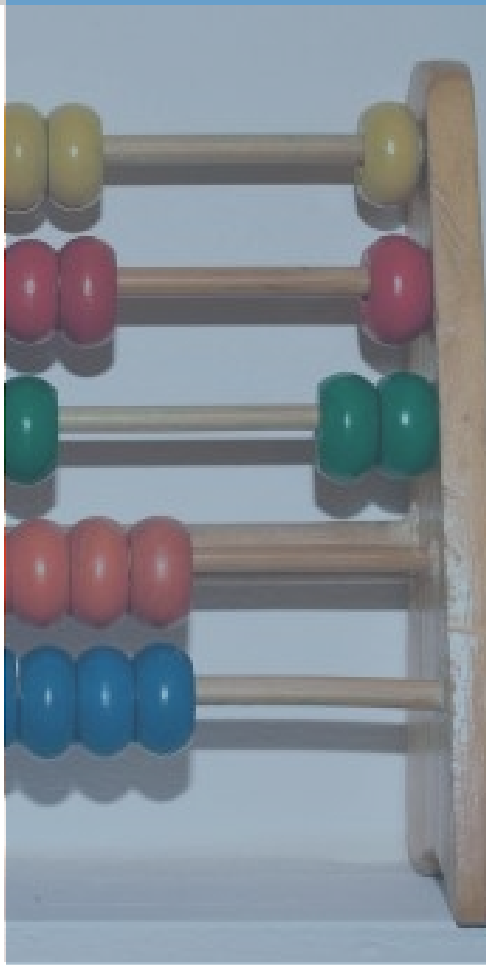


Grundlegende statistische Verfahren für archäologische Datenanalyse in R

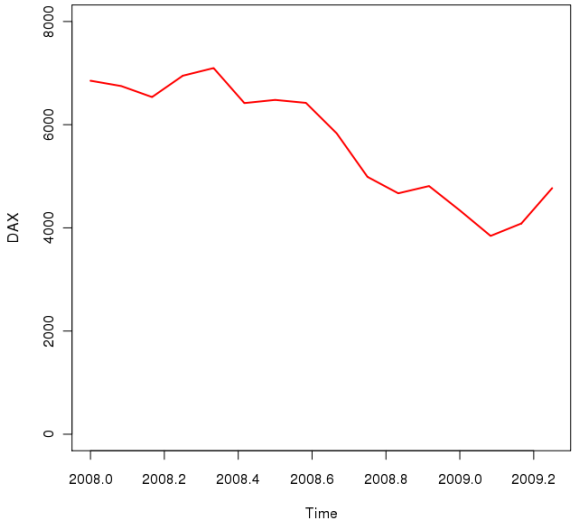
Christian-Albrechts-Universität zu Kiel

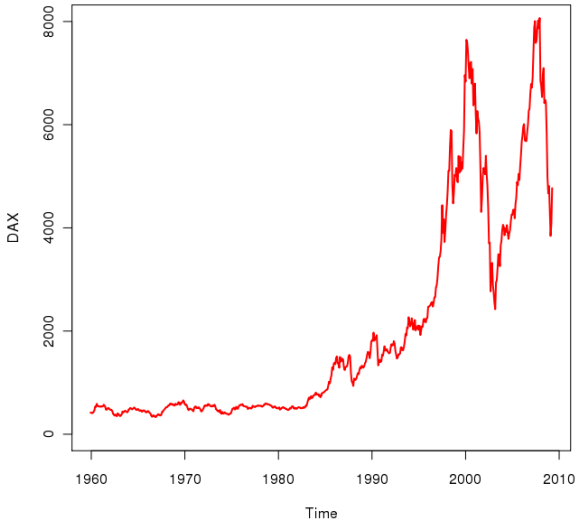
Gestalten von Diagrammen

Immer schön ehrlich bleiben!
Wahl der Darstellung hat starken Einfluß auf die produzierte Aussage.



Grundlegende statistische Verfahren für archäologische Datenanalyse in R





C A U
Christian-Albrechts-Universität zu Kiel

Gestalten von Diagrammen

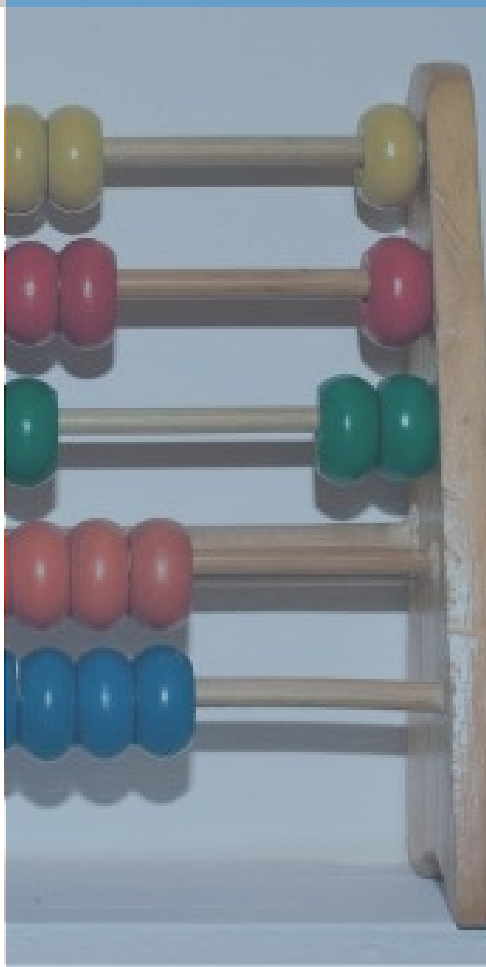
Immer schön ehrlich bleiben!
Wahl der Darstellung hat starken Einfluß auf die produzierte Aussage.

Klare Darstellung!
Anteil der Tinte je dargestellter Information minimieren

Passendes Diagramm für die Daten wählen!
Nominal-Ordinal-Intervall-Verhältnisskala beachten

Was ist darzustellen?	*besonders gut geeignet*	*ungeeignet*
Einheiten vom Ganzen: wenige Teile	Kreisdiagramm, Stapelbalkendiagramm	
Einheiten vom Ganzen: viele Teile	Stapelbalkendiagramm	
Mehrfachantworten	horizontales Balkendiagramm	Kreisdiagramm, Stapelbalkendiagramm
Vergleich verschiedener Ausprägungen mehrerer Variablen	Gruppiertes Balkendiagramm	
Vergleich verschiedener Anteile vom Ganzen	Stapelbalkendiagramm	
Vergleich von Entwicklungen	Liniendiagramm	
Häufigkeitsverteilung einer Variablen	Histogramm	
Übereinstimmung zweier Variablen	Streudiagramm	

http://www.univie.ac.at/ksa/elearning/cp/quantitative/quantitative-123.html



Nächste Sitzung 18. November 2010: Deskriptive Statistik

Weiterführendes Material unter

http://www.statmethods.net/
http://de.wikibooks.org/wiki/GNU_R:_Diagramme
http://docs.ggplot2.org/current/
http://rgraphgallery.blogspot.de/