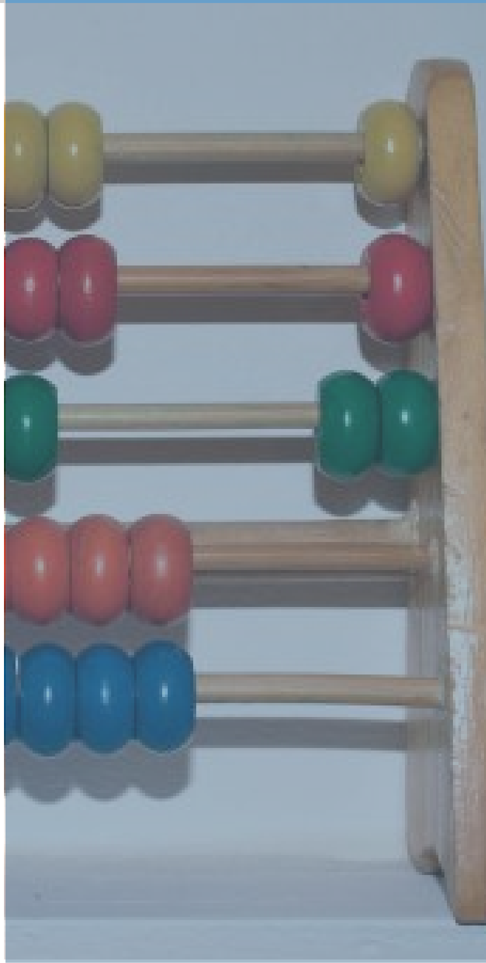
Basic statistic techniques for (archaeological) data analysis in R

Christian-Albrechts-Universität zu Kiel

03_explorative_statistics-graphical_display

Tables and charts

Loading data for the following steps

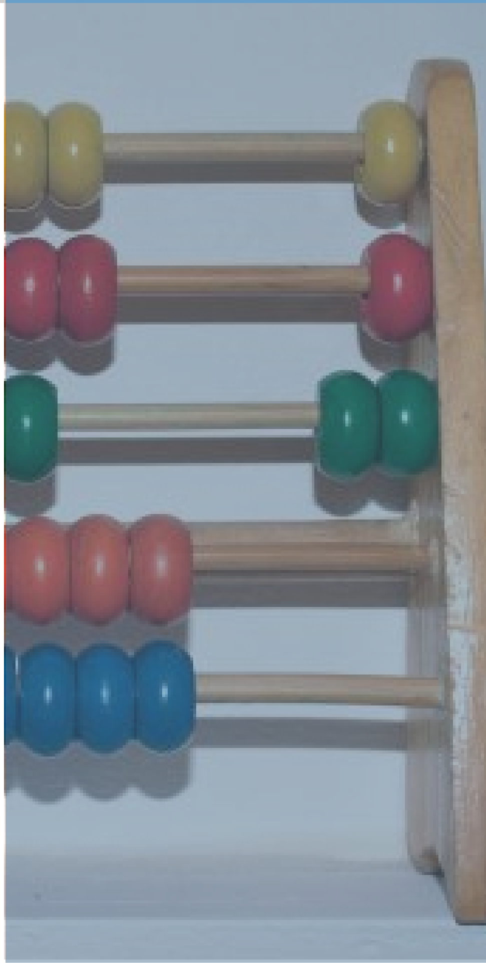


Read the data of the Kursteilnehmer:

```
> setwd("--your R-directory--")
> 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
```

Cross tables (contingency tables)



For summary of data:

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

                    Afrika Asien Europa Mittelamerika Südamerika
  Annalena Bock          0     0      3             0          0
  Henry Skorna           0     2      1             0          0
  Janna Kordowski        0     0      3             0          0
  Saryn Schlotfeldt      0     0      0             2          1
  Timo von Holtz        13     0      0             0          0
> addmargins(tabelle)

                    Afrika Asien Europa Mittelamerika Südamerika Sum
  Annalena Bock          0     0      3             0          0   3
  Henry Skorna           0     2      1             0          0   3
  Janna Kordowski        0     0      3             0          0   3
  Saryn Schlotfeldt      0     0      0             2          1   3
  Timo von Holtz        13     0      0             0          0  13
  Sum                   13     2      7             2          1  25
```

Basics about charts

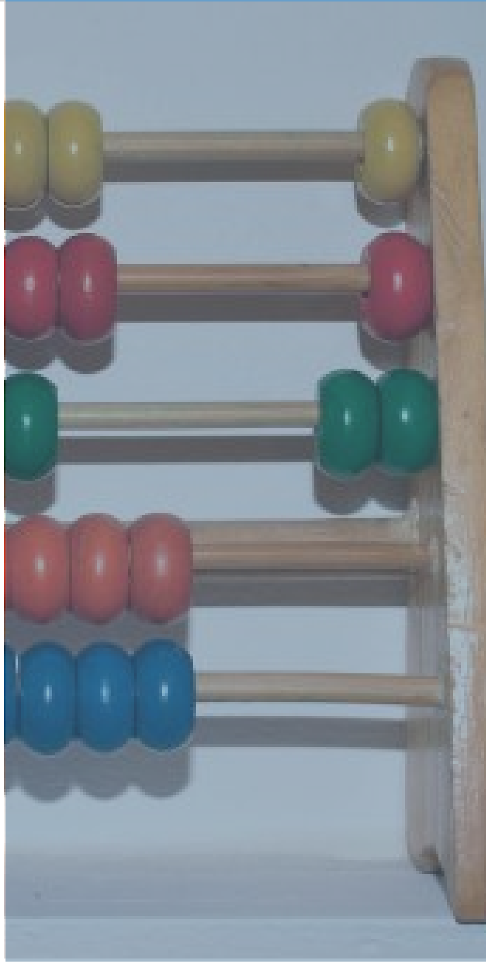
Principles for good charts according to E. Tufte:
(The Visual Display of Quantitative Information. Cheshire/
Connecticut: Graphics Press, 1983)

- „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.”

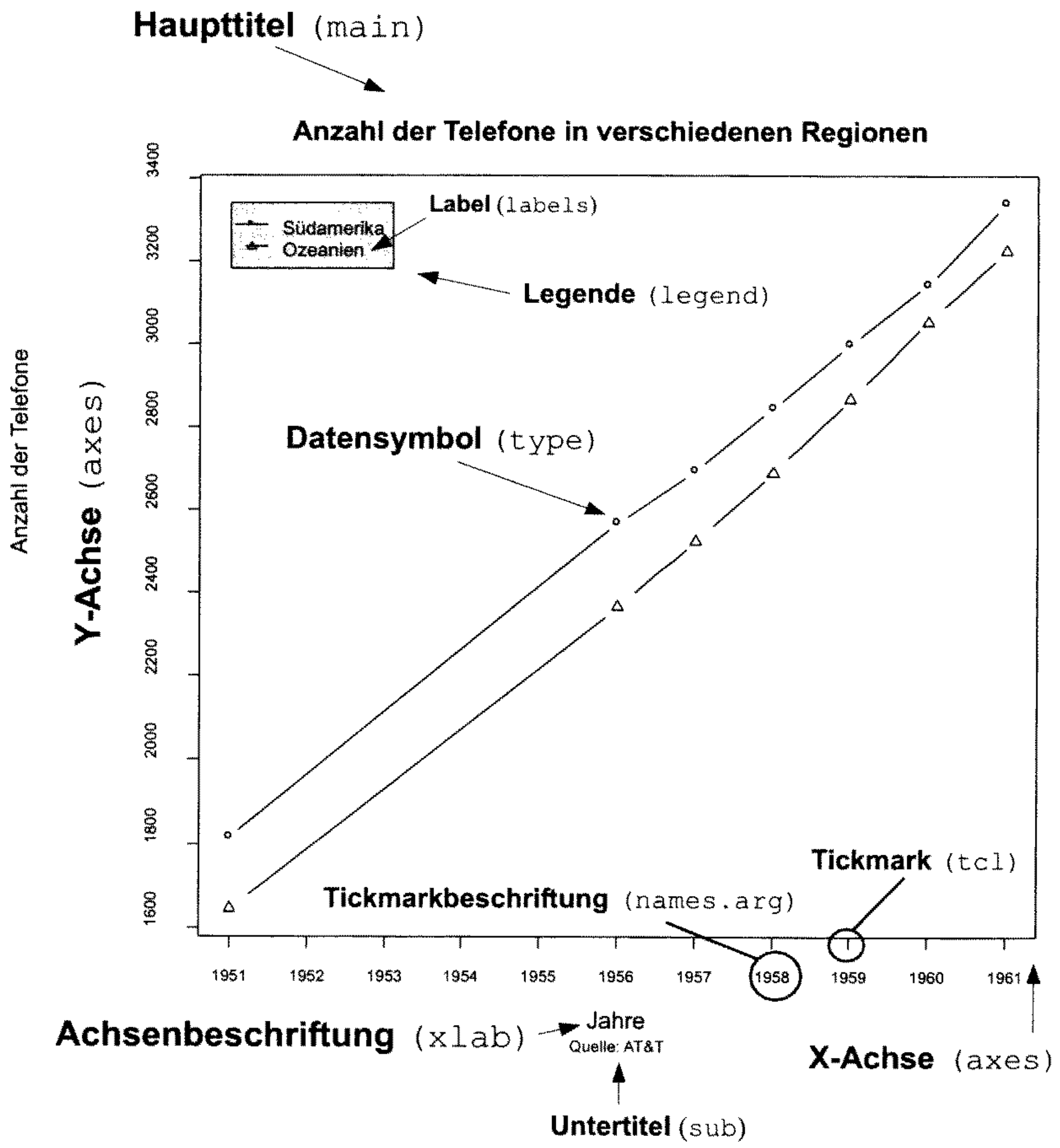
- Data-ink ratio = „proportion of a graphic’s ink devoted to the non-redundant display of data-information“ (kein chartjunk!)

- „Graphical excellence is often found in simplicity of design and complexity of data.“

after Müller-Scheeßel



Basic statistic techniques for (archaeological) data analysis in R



Plot [1]

Basic drawing function of R:

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

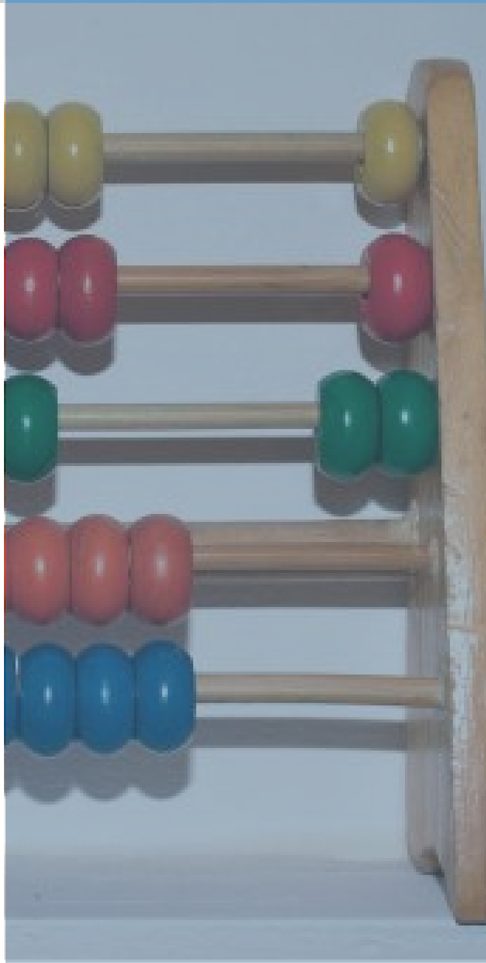
options:

p	– points (default)
l	– solid line
b	– line with points for the values
c	– line with gaps for the values
o	– solid line with points for the values
h	– vertical lines up to the values
s	– stepped line from value to value
n	– empty coordinate system

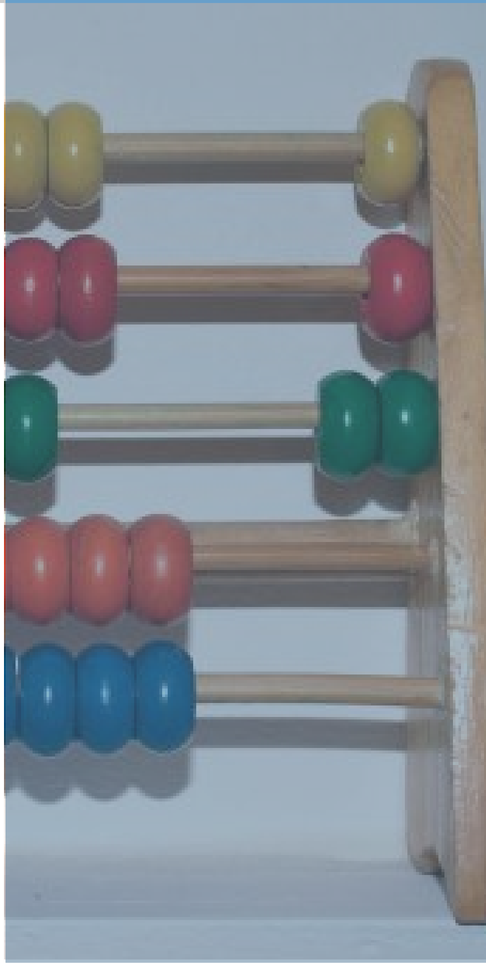
```
> plot(laender$Einwohnerzahl,type="b")
```

Intelligent system: automatic determination of variable type, drawing of the appropriate chart

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



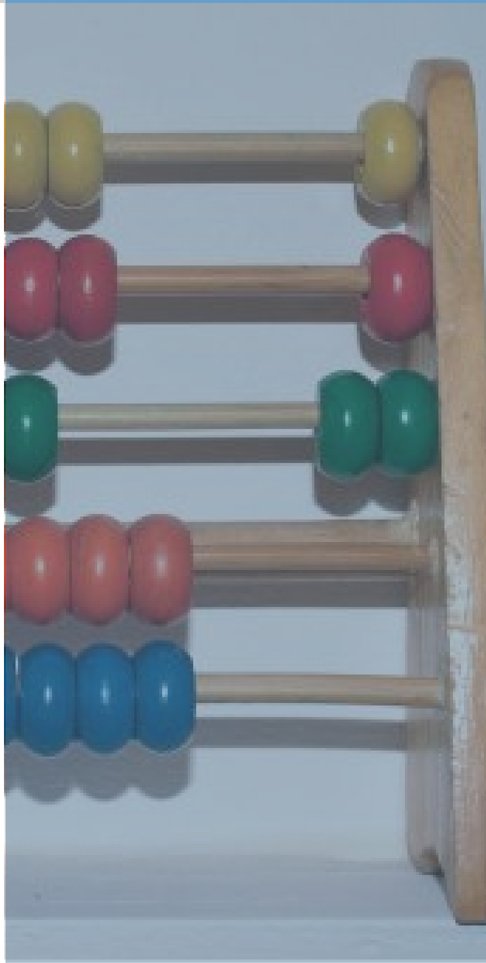
Plot [2]



Optional components and text:

```
> plot(laender$Fläche,laender$Weltrang.CPI,
     xlim=c(0,2500000), # limits of the x axis
     ylim = c(0,200), # limits of the y axis
     ylab = "rank according to CPI", # label of the x axis
     xlab = "area", # label of the y axis
     main = "area vs. rank according to BPI", # main title
     sub="example plot" # subtitle
)
```

Plot [3]



Plot do a lot for you:

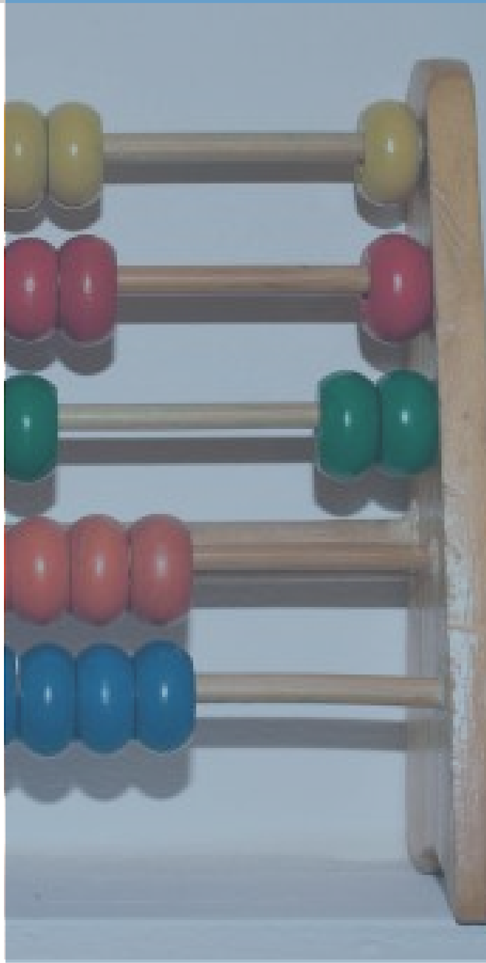
Opens a window for display
Determines the optimal size of the frame of reference
Draws the coordinate system
Draws the values

Gives a „handle“ back for further additions to the plot, e.g.:

lines	– additional lines to an existing plot
points	– additional points to an existing plot
abline	– additional special lines to an existing plot
text	– additional text on choosen position to an existing plot

Additional possiblities for “decorations”: ? par

Plot [4]



Add additional elements:

Drawing lines

```
> 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))
```

Drawing text

```
> text(2000000, mean(laender$Weltrang.CPI), # position at x 20 und y mean
label = paste("MW (CPI)= ", # text is concatenate via paste
round(mean(laender$Weltrang.CPI,na.rm=T))),
pos = 3, # position above
cex = 0.7 # font size 70%
)
```

Export the graphics

With the GUI:
File → Save as...

With the commando line:
As vector file

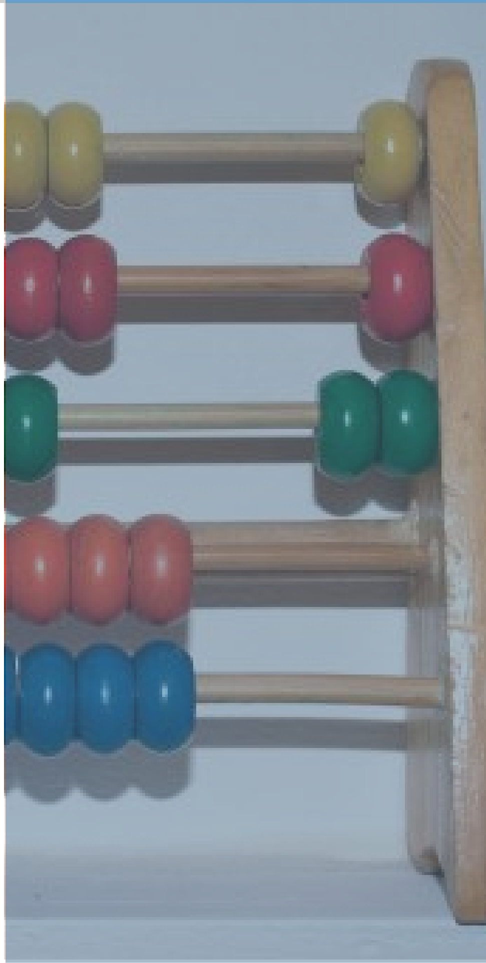
```
> dev.copy2eps(file="test.eps")
> dev.copy2pdf(file="test.pdf")
```

As bitmap file

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

Possible are “png”, “jpeg”, “tiff”, “bmp”

SavePlot can save sometimes also vector files (dependent on operation system and installation)



Pie chart [1]

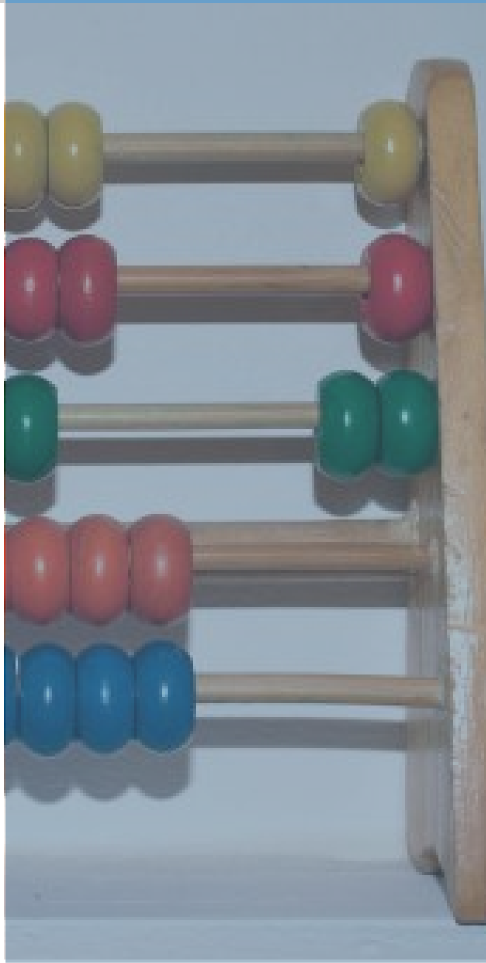
The classical one – but also with R not much better...
Used to display proportions, suitable for nominal data

$$a_i = \frac{h_i}{N} \cdot 360°$$

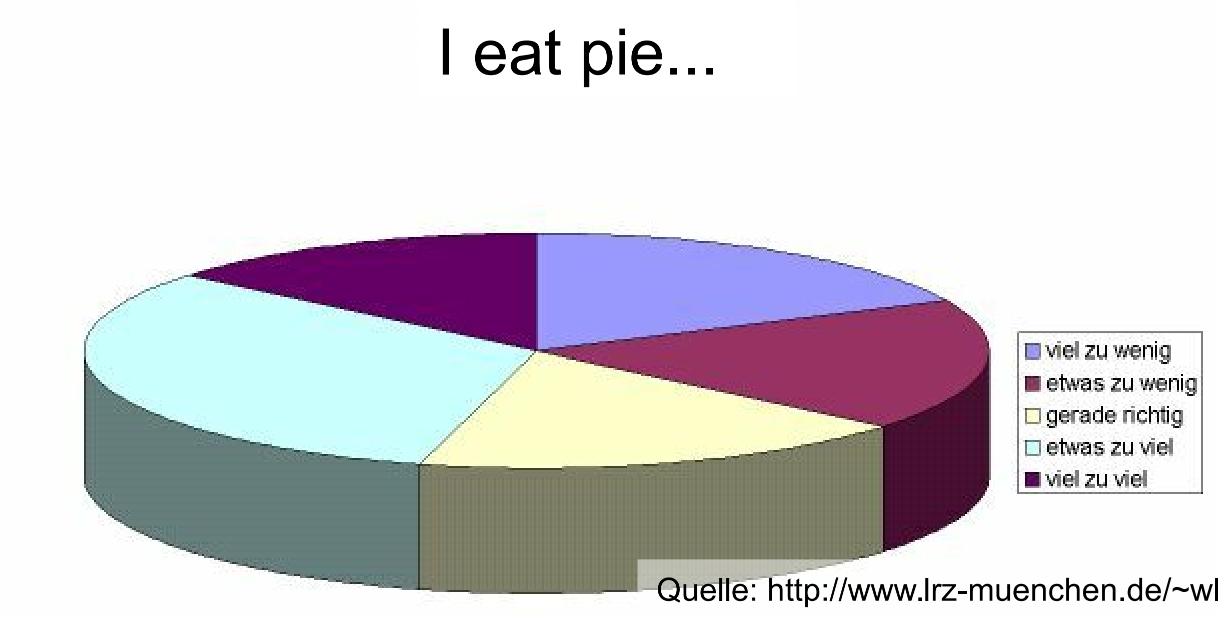
Disadvantages:

Color selection can influence the perception (red is seen larger then gray)
Small differences are not easy visible

totally No-Go: 3d-pies!!!

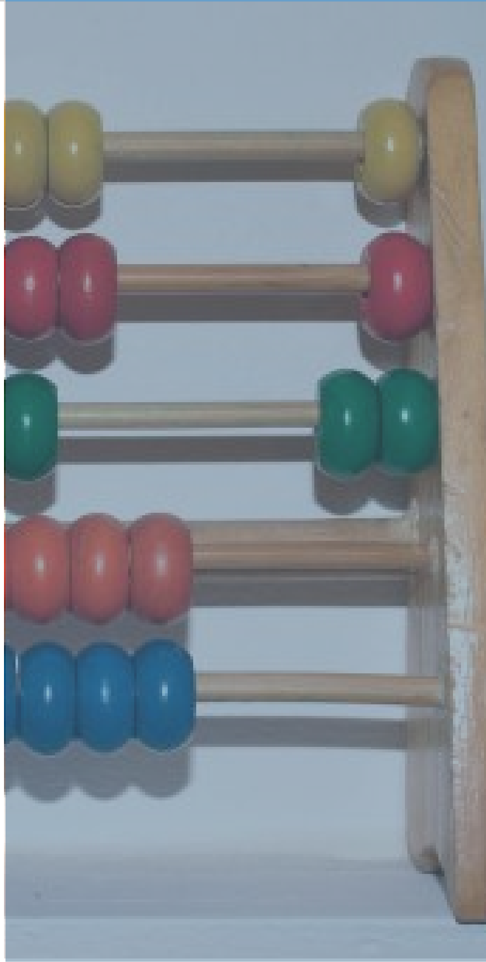


Pie chart [2]

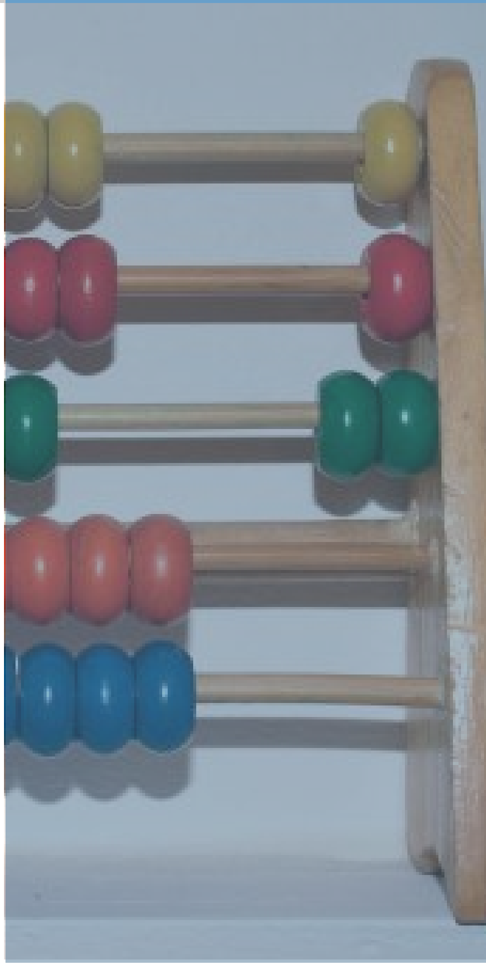


Quelle: http://www.lrz-muenchen.de/~wlm

The pieces »viel zu wenig«, »etwas zu wenig« und »gerade richtig« have exactly the same size, the piece »viel zu viel« is a bit smaller.



Pie chart [3]



Pies in R

Data are a vector of counts

```
> table(laender$kontinent)

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

Color palette:

The standard palette is pastel, if you prefer another:

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

Bar plot [1]

Generally the better alternative...

Bar plots are suitable for display of proportions as well as for absolute data. They can be used for every level of measurment.

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

With names:

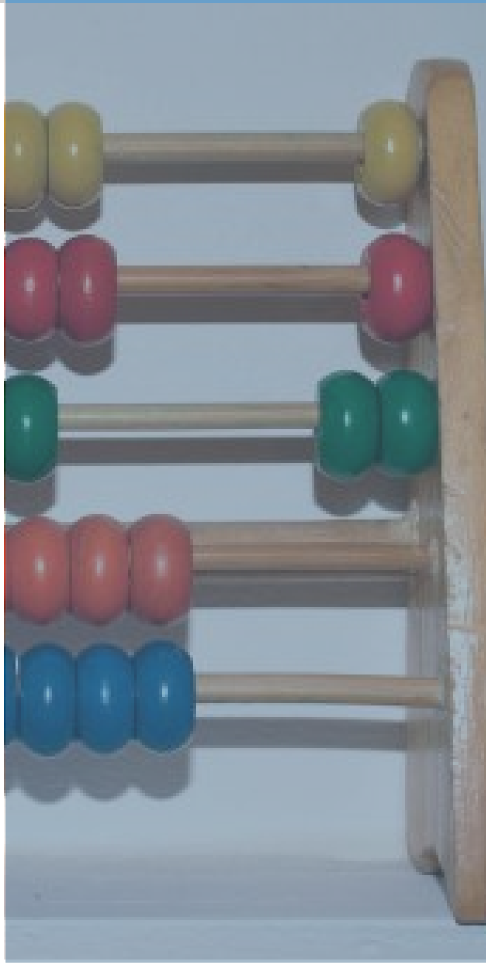
```
> par(las=2)
> barplot(laender$Fläche.in.km.,names.arg=laender$Name)
```

With title:

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

Horizontal:

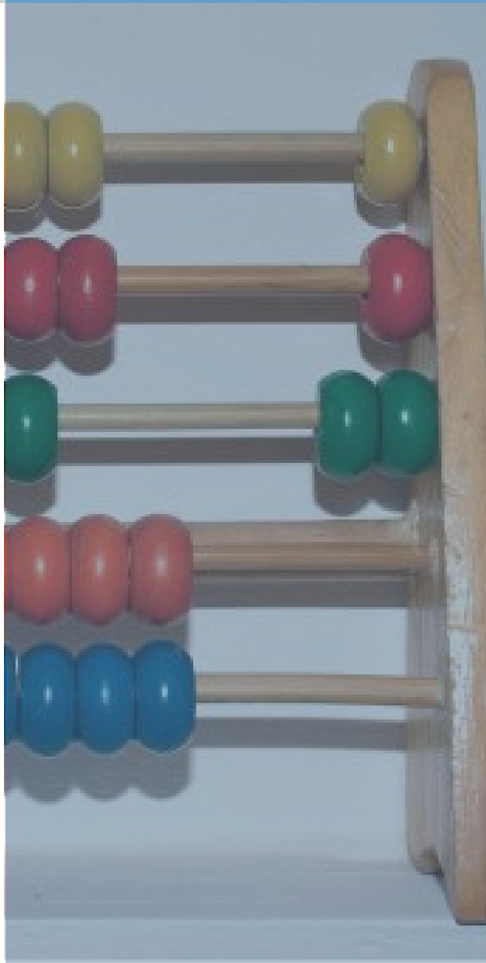
```
> barplot(table(laender$kontinent),horiz=T,cex.names=0.5)
```



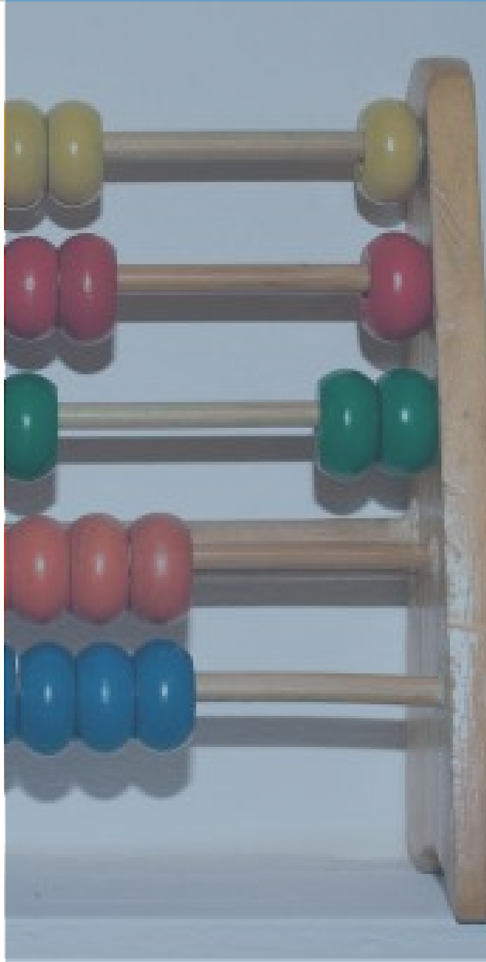
Bar plot [2]

Display of counts

```
> 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))
```



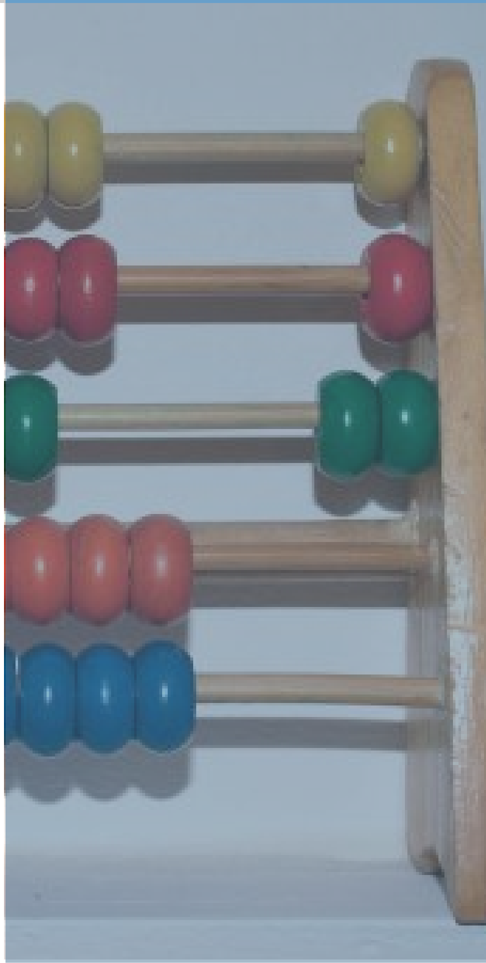
Bar plot [3]



Display of proportions

```
> 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("ratio of contributers \n by continent", outer=TRUE, line=-
3)
```

Bar plot [4]



Problems with bar plots – and also with many other charts

Percent vs. count: percents often distort the relations

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

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

Scales: the choosen limits of the axes can distort the relations

```
> 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))
```

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

One of the best (my favorite)!

Used to display the distribution of values in a data vector of metrical (interval, ratio) scale

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

> boxplot(1:9)
```

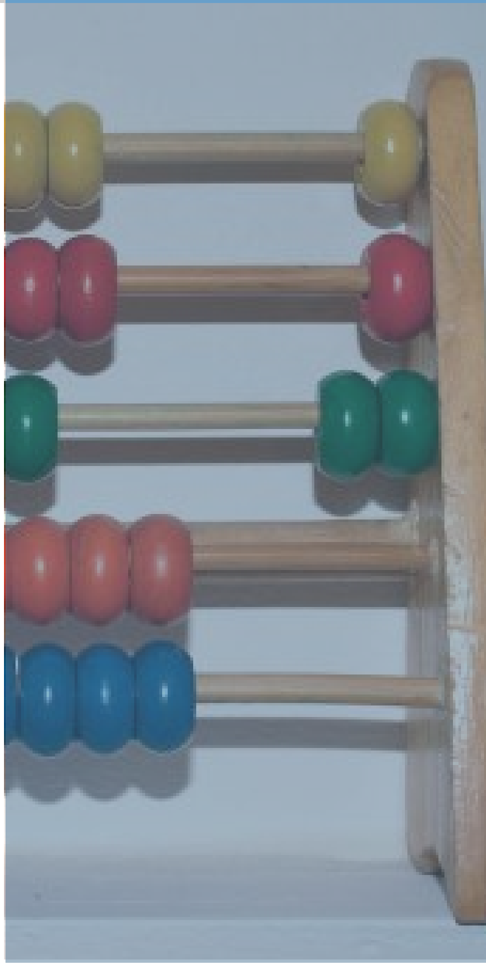
Box: the inner both quantiles

Whisker: last value < than 1.5 times the distance of the inner quantile

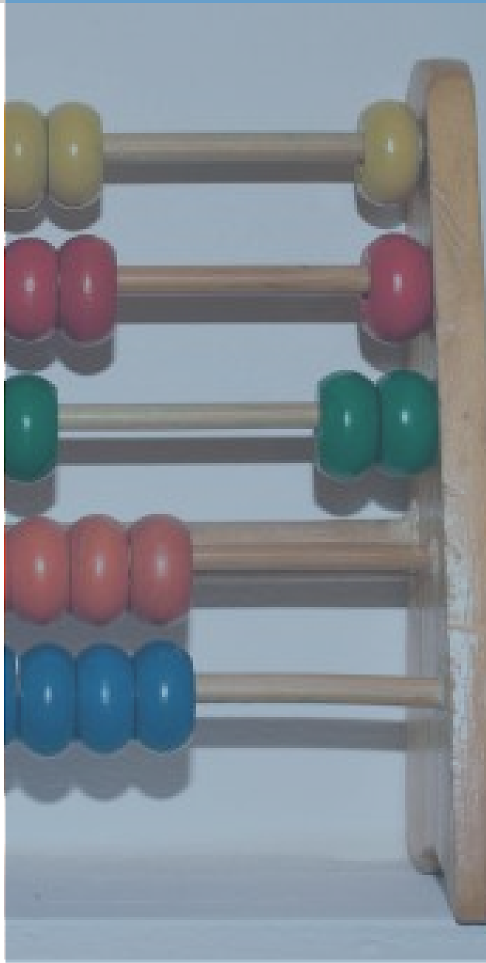
```
> boxplot(laender$Fläche)
> boxplot(laender$Fläche.in.km.~laender$einlieferer)
```

More beautiful:

```
> 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")
```



scatterplot



For 2 discrete variables

Used to display a variable in relation to another one. Generally for all scales suitable, but for nominal and ordinal scale other charts are often better.

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

Call additional libraries:

```
> library(car) # library for regression analysis
> scatterplot(Fläche.in.km.~Weltrang.CPI,data=laender)

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

Histogramm

Used for classified display of distributions

Data reduction vs. precision: Display of count values of classes of values

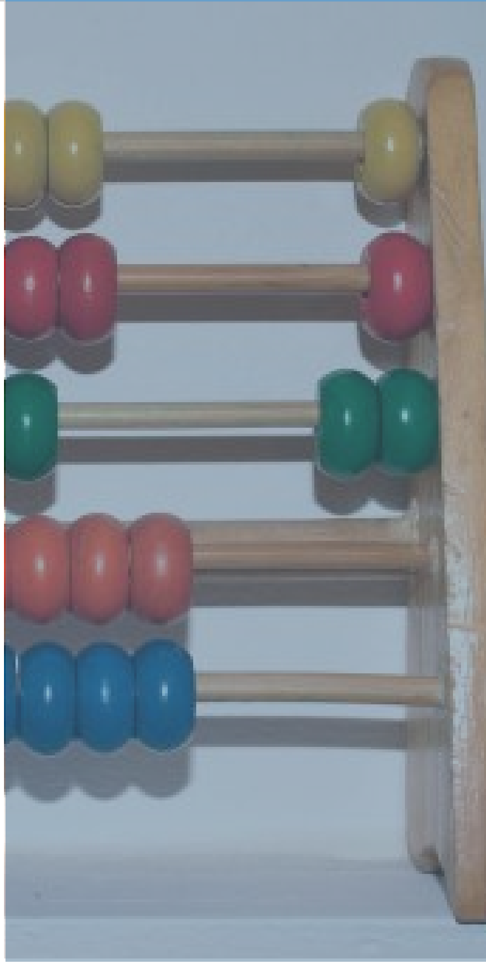
```
> hist(laender$Fläche)
> hist(laender$Fläche, labels=T)
> hist(laender$Fläche, labels=T, breaks=20)
>
```

More beautiful

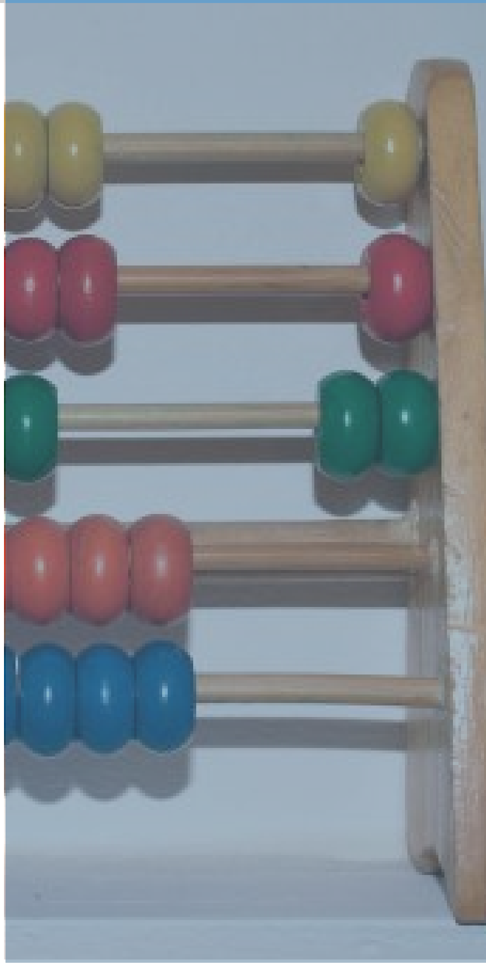
```
> hist(laender$Fläche,breaks=20,labels=T, col="red",
xlab="Fläche", main="histogram of area of selected
countries")
```

Disadvantages:

Data reduction vs. precision → loss of information
Actual display depends strongly on the choosen class width



steam-and-leaf chart



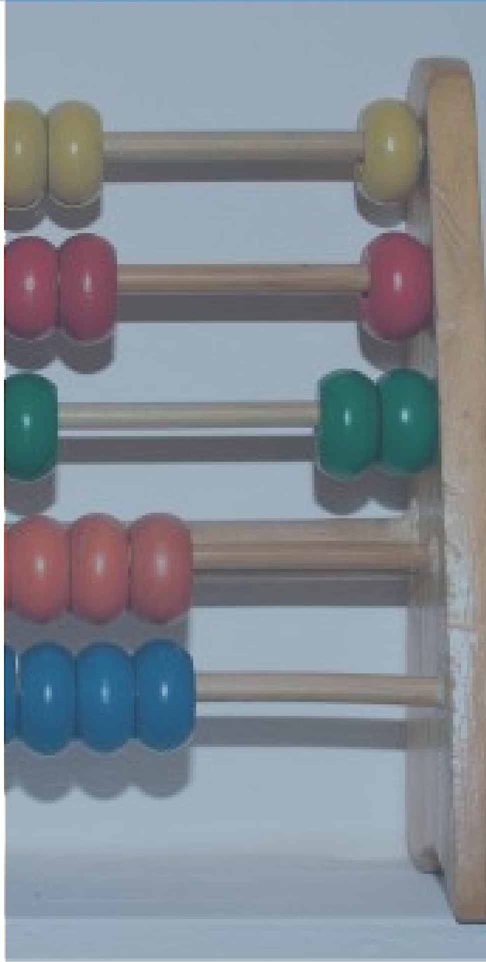
An attempt to overcome the disadvantages of a histogram

Is not very often used. Scales like histograms.

```
> 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
```

Advantage: Information about the distribution inside the classes and the absolute values are (partly) visible.

kernel smoothing (kernel density estimation)



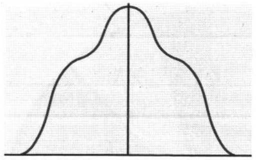
Another attempt to overcome the disadvantages of a histogram
The distribution of the values is
Die Verteilung der Werte wird considered and a distribution curve is calculated. Continuous distributions are better displayed, without artificial breaks. Scales like histograms.

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

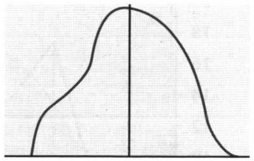
Histogram and kernel-density-plot together

```
> 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)
```

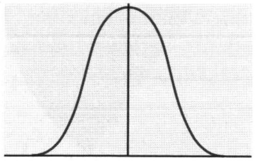
Shapes of distributions (after Bortz 2006)



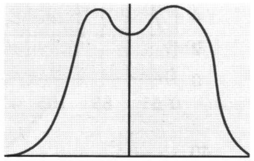
symmetric



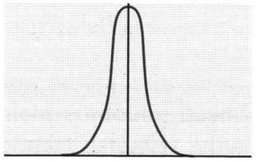
asymmetric



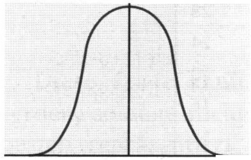
unimodal



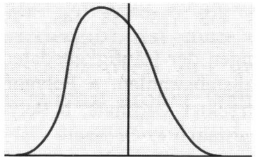
bimodal



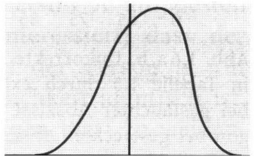
? e schmalgipflig



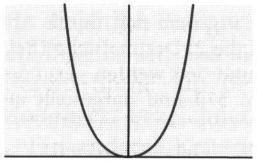
f breitgipflig ?



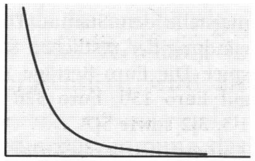
Positive skewed



negative skewed

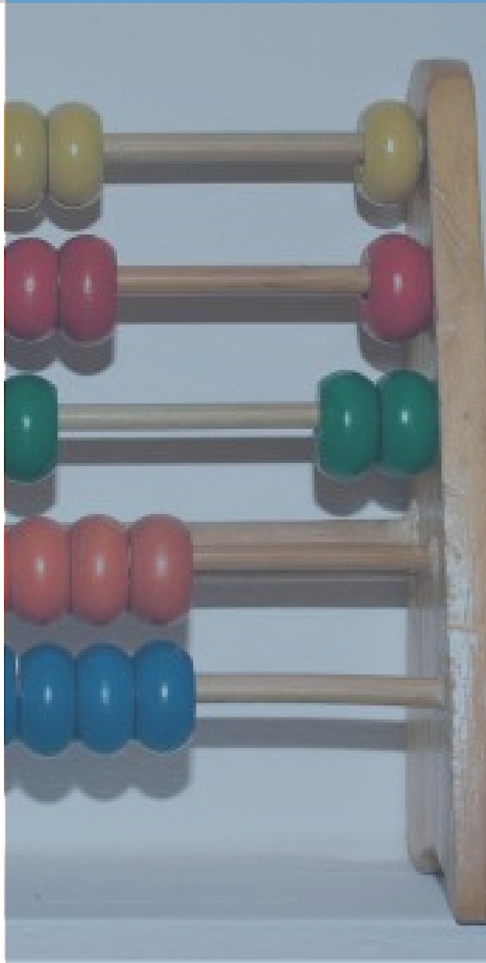


U-shaped



falling

Culmulative frequency distribution



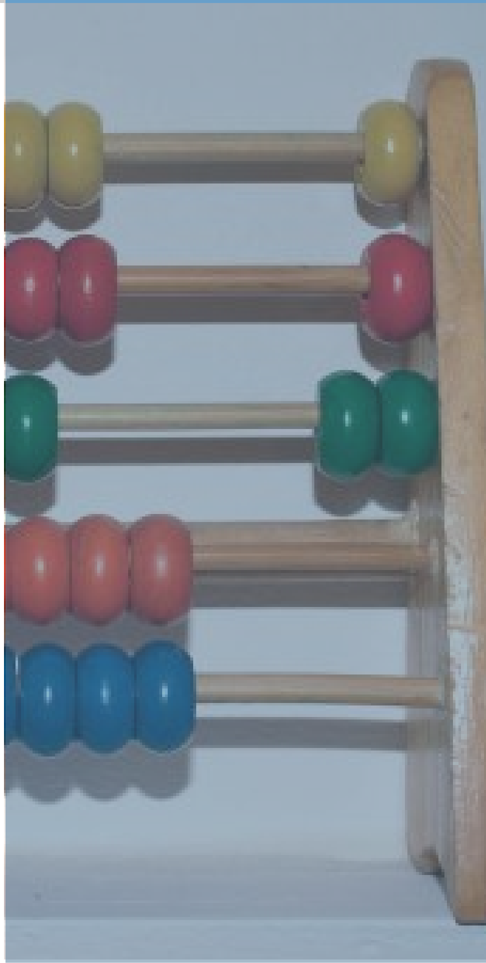
Display of the proportions of ordinal variables

Example from Shennan: Counts of burials by age

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("cumulative ratio of burried by age class")
```

Triplot

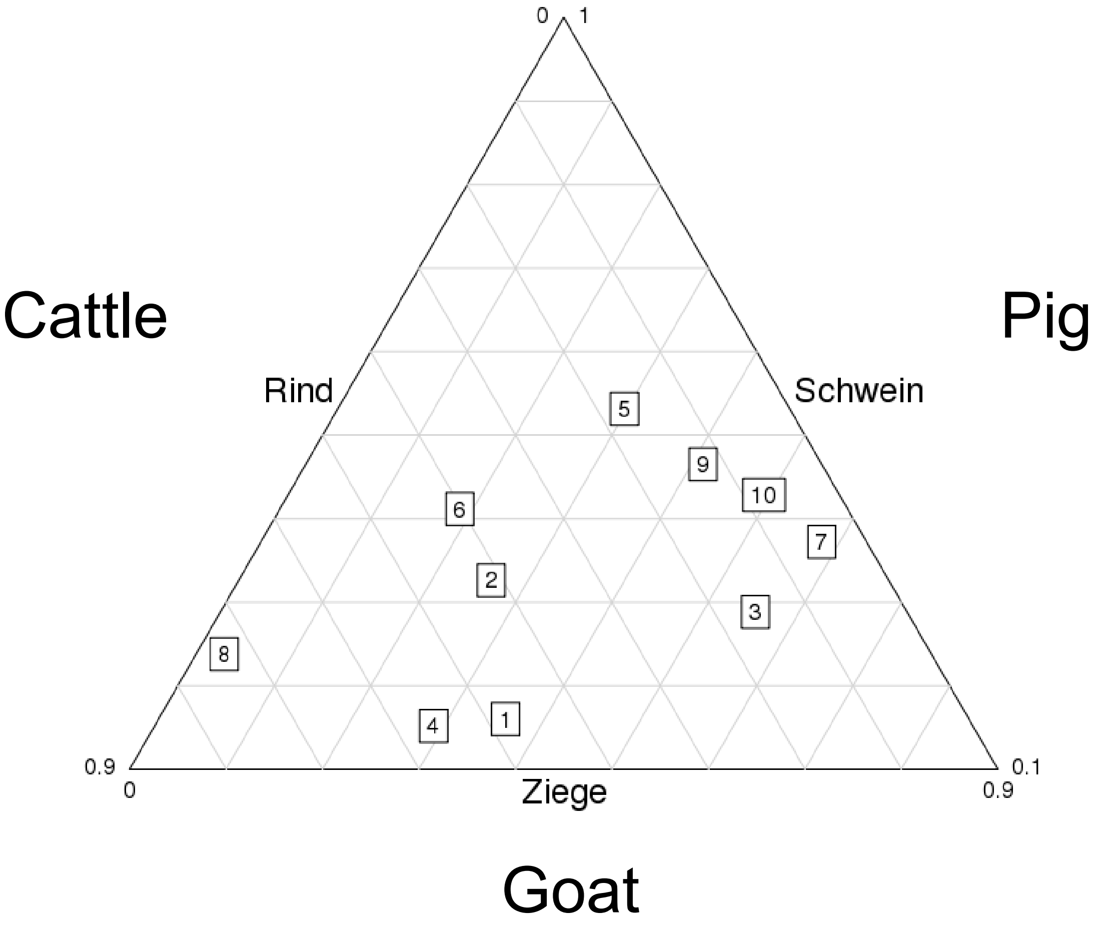


Simplest kind of multivariate chart

Used for display of the proportions of 3 exclusive Variables
General suitable for all levels of measurement. Datas are converted into percent.

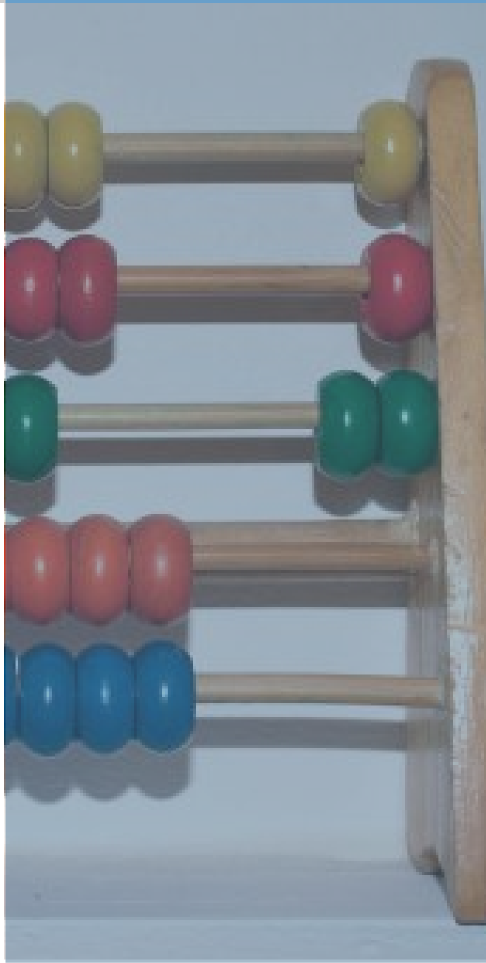
```
> library(ade3)
> test<-matrix(round(abs(rnorm(30)*100)),ncol=3)
> colnames(test)<-c("cattle","goat","pig")
> test
      cattle  goat pig
 [1,]    195   146  65
 [2,]     96    61  76
 [3,]     36   127  66
 [4,]    114    59  31
 [5,]     49    85 152
 [6,]    168    78 172
 [7,]     10   125  80
 [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)
```

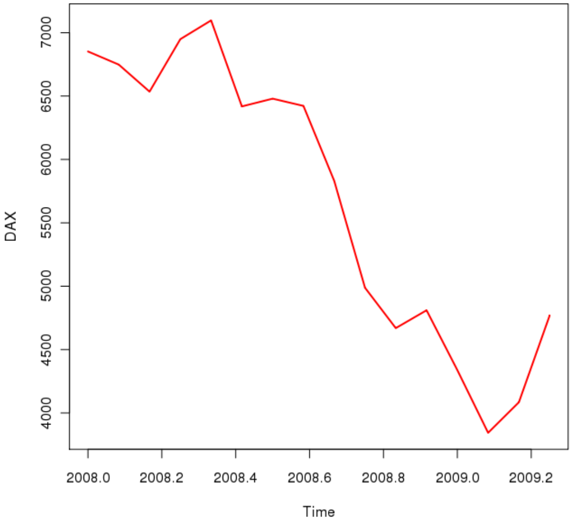
Simulated triplot of the proportions of animal bones from different settlements

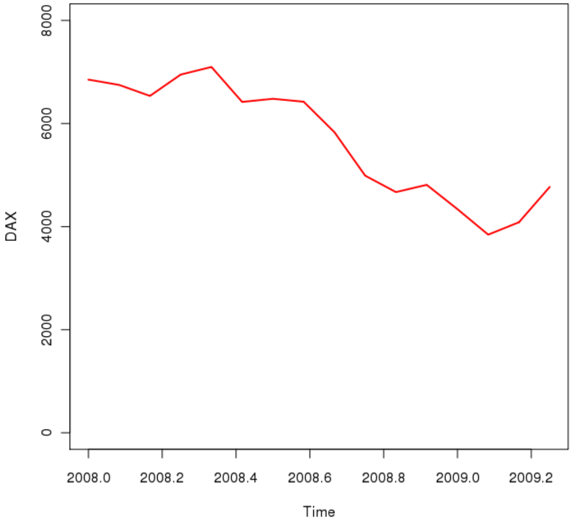


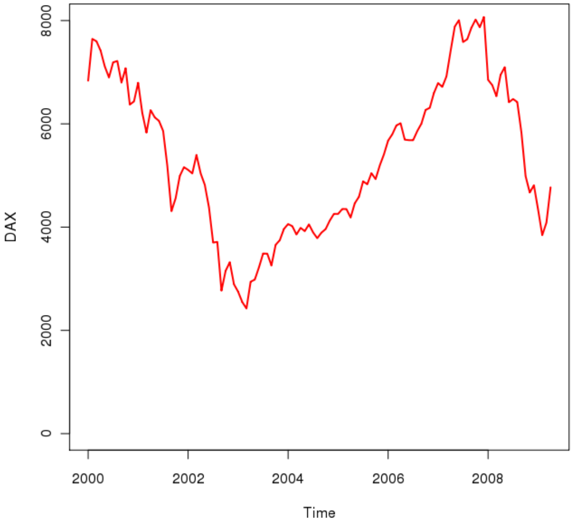
Style of charts

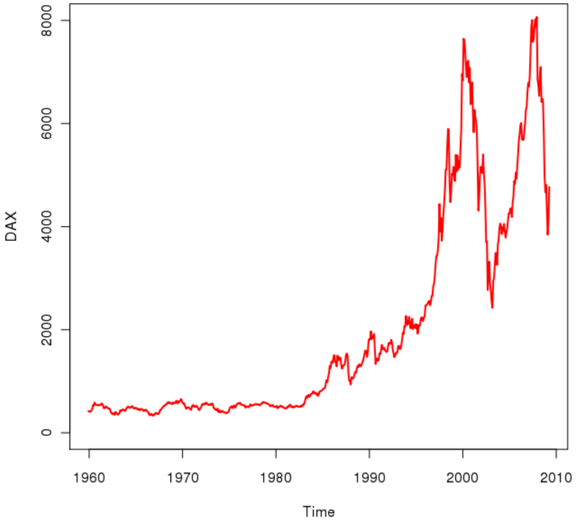
Stay honest!
Choice of display has a strong influence on the statement.




DAX
4000
4500
5000
5500
6000
6500
7000
2008.0
2008.2
2008.4
2008.6
2008.8
2009.0
2009.2
Time


DAX
0
2000
4000
6000
8000
2008.0
2008.2
2008.4
2008.6
2008.8
2009.0
2009.2
Time


DAX
0
2000
4000
6000
8000
2000
2002
2004
2006
2008
Time


DAX
0
2000
4000
6000
8000
1960
1970
1980
1990
2000
2010
Time

Style of charts

Stay honest!
Choice of display has a strong influence on the statement.

Clear layout!
Minimise Ratio of ink per shown information!

Use the suitable chart for the data!
Consider nominal-ordinal-interval-ratio scale

What to display	suitable	Not suitable
Parts of a whole: few	Pie chart, stacked bar plot	
Parts of a whole: few	Stacked bar plot	
Multiple answers (ties)	Horizontal bar plot	Pie chart, stacked bar plot
Comparison of different values of different variables	Grouped bar plot	
Comparison of parts of a whole	Stacked bar plot	
Comparison of developments	Line chart	
Frequency distribution	Histogram, kernel density plot	
Correlation of two variables	scatterplot