Christian-Albrechts-Universität zu Kiel

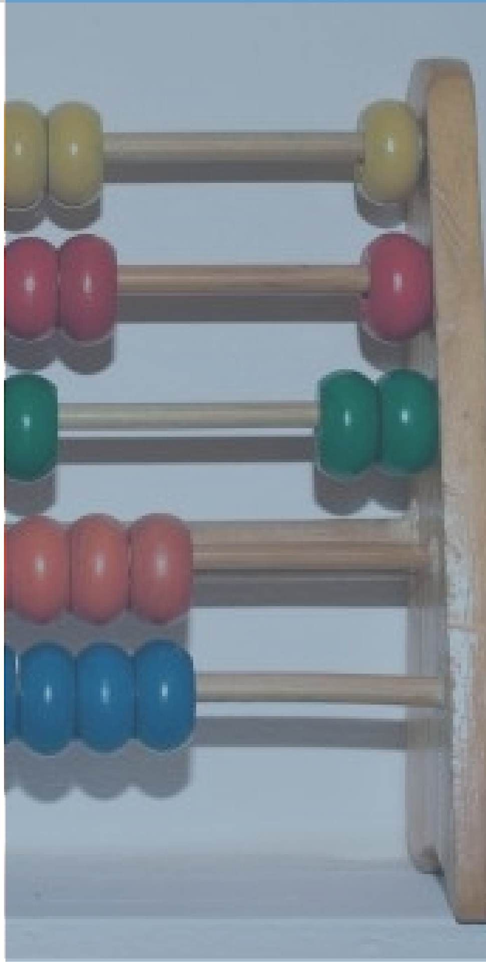
10_regression_und_korrelation

Regression, Korrelationskoeffizient, Kendalls Tau

Laden der Daten für weitere Schritte

Einlesen der Daten der Kursteilnehmer:

```
> setwd("--ihr R-Verzeichnis--")
> kursdaten<-read.csv2("kursdaten.csv")
> kursdaten
  Schuhgroesse Geschlecht  Gr
1           40          w 174
2           40          w 163
3           43          m 182
4           44          m 175
5           43          m 173
6           49          m 198
7           44          m 179
8           37          w 163
9           42          m 181
```



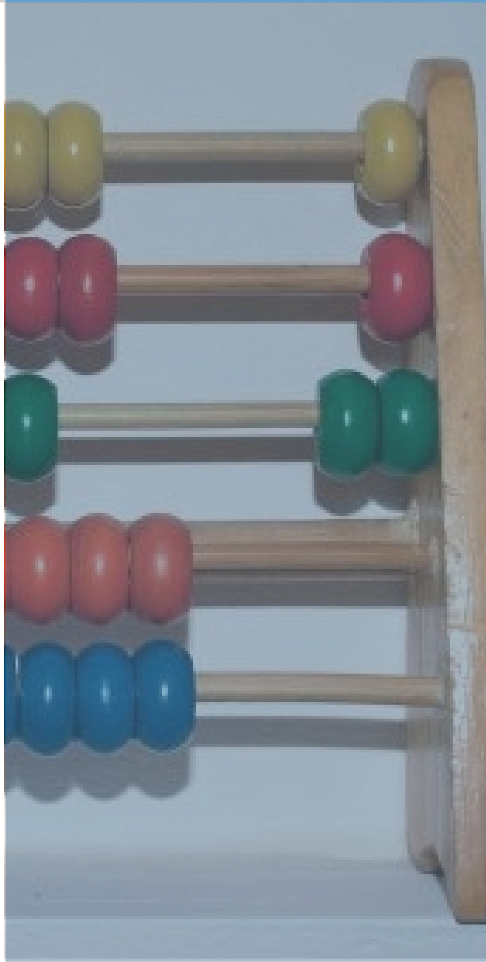
Scatterplot

Darstellung von zwei Variablen zueinandern

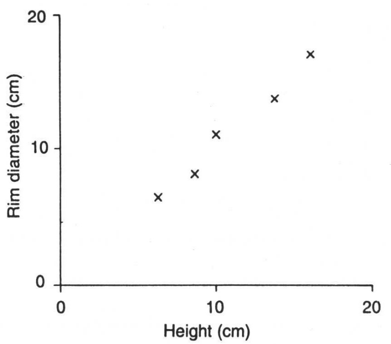
```
> plot(kursdaten$Gr,kursdaten$Schuhgroesse)
```

Sichtbar: Vergrößert sich die eine, vergrößert sich die andere Variable.

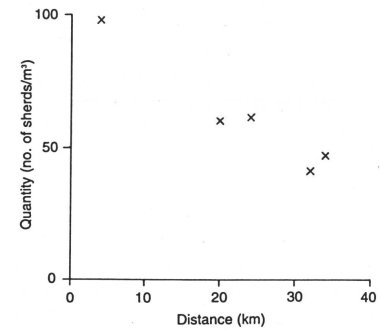
Weitere ablesbare Eigenschaften:
Richtung der Beziehung (größer-> größer vs. größer -> kleiner)
Linearität der Beziehung (monoton, nicht monoton
Stärke/Strenge der Beziehung (punkte nahe vs. weit entfernt von einer gedachten Linie)



Grundlegende statistische Verfahren für archäologische Datenanalyse in R

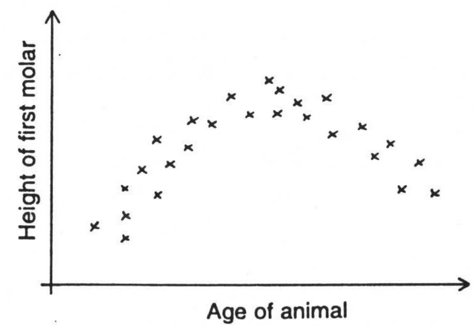


Positive Regression

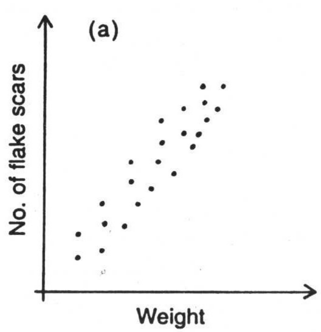


Negative Regression

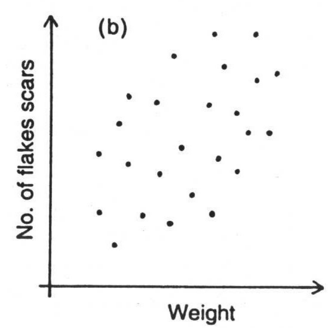
Nichtlineare
monotone Regression



Nichtlineare
nichtmonotone Regression



Starke Korrelation



Schwache Korrelation

Richtung und Linearität von Beziehungen

Richtung

Gibt an, ob eine Variable mit der anderen steigt (positiv) oder fällt (negativ)

Variablen: mögliche Ursache (unabhängige Variable) und interessierende Wirkung (abhängige Variable)

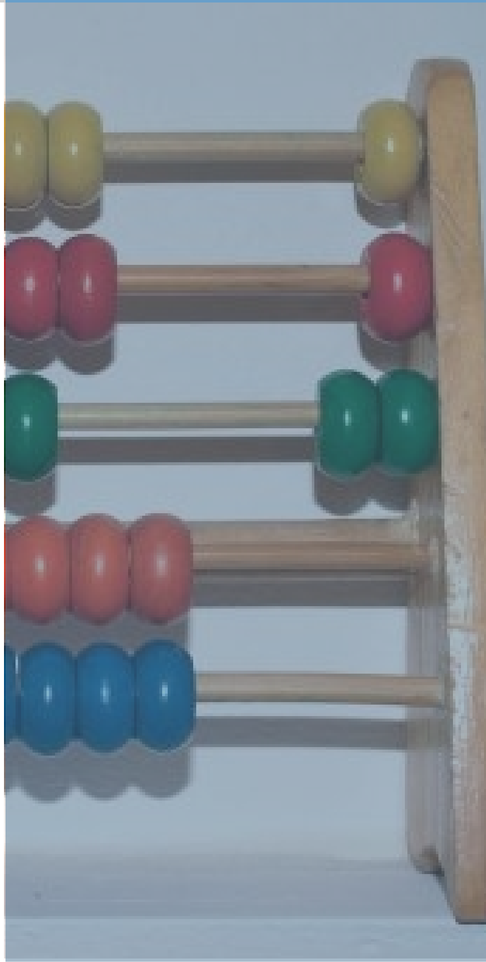
Linearität

Es gibt lineare und nicht-lineare Regressionen

Nicht-lineare Regressionen, mögliche Ursachen:

Kombination von verschiedenen (linearen?) Einflüssen: multiple Regressionsanalyse

Einflußfaktor wirkt sich nicht linear aus: nichtlineares Modell (Quadratisch oder höheres Polynom, Schwellenwertsysteme etc.)



Regression: Formel

Was wir noch aus dem Schulunterricht wissen...

Die Formel für eine lineare Gleichung setzt sich zusammen aus einer Steigung (b) und einer Inhomogenität (Verschiebungskonstante a)

$$y = a + bx$$

$$b = \frac{(y_2 - y_1)}{(x_2 - x_1)} \quad a = y_1 - b * x_1$$

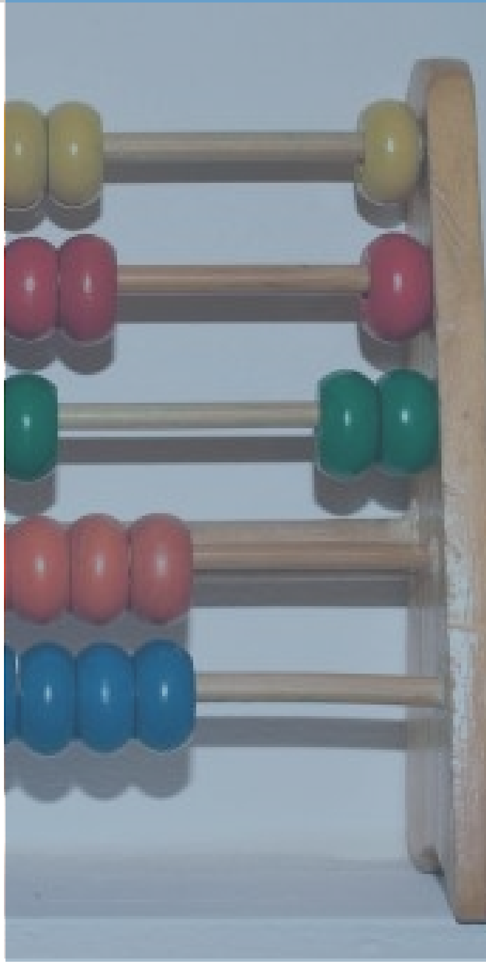
Beispiel: $\{1,3\}, \{2,5\}, \{3,7\}$ ...

$$b = \frac{(5-3)}{(2-1)} = 2$$

$$a = 3 - 2 * 1 = 1$$

$$y = 1 + 2 * x$$

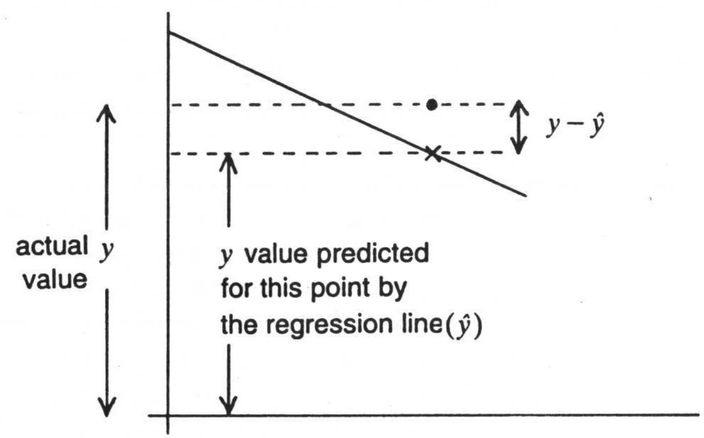
Aber: das funktioniert nur bei perfekter Korrelation, was bei abweichenden (statistischen) Werten?



Regression: least-squares Methode (Methode der kleinsten Quadrate) [1]

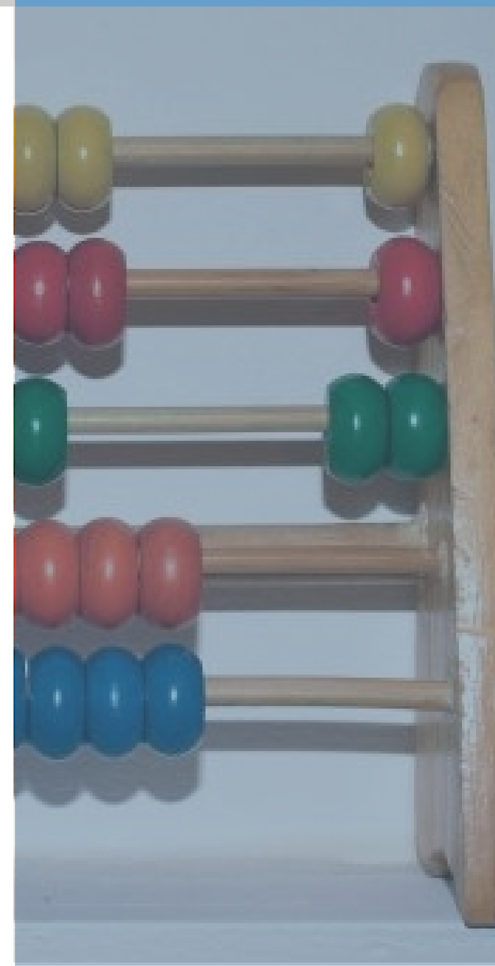
Schätzung der optimalen Annäherung mit der least-square Methode

Für Werte, die nicht genau einer Geraden entsprechen, muss eine optimale Annäherung gefunden werden.

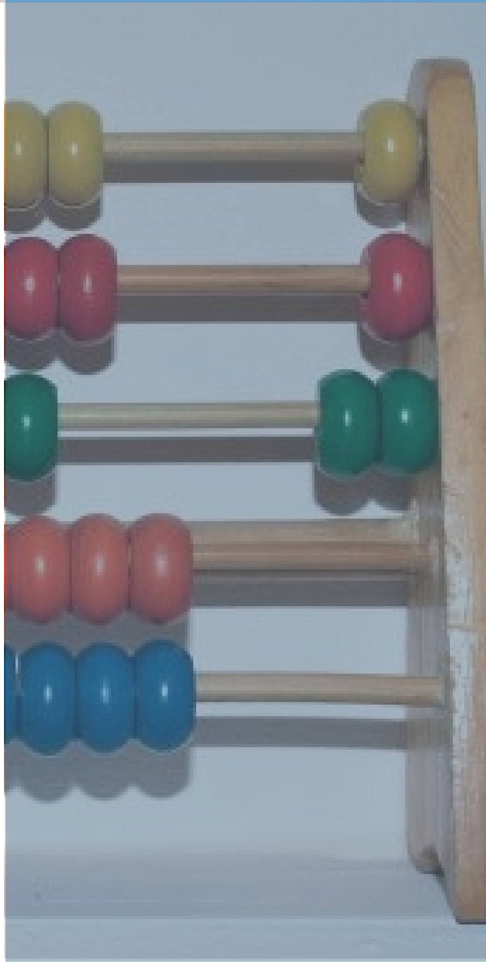


Der absolute Abstand zwischen dem echten y-Wert und dem geschätzten y-Wert soll möglichst klein sein, es gilt:

$$min \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$



Regression: least-squares Methode (Methode der kleinsten Quadrate) [2]



Formel für die least-square Methode

$$min \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = b_{min} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$a_{min} = \bar{y} - b * \bar{x}$$

für Körpergröße gg. Schuhgröße: 174,40 163,40 182,43 175,44 173,43 198,49 179,44 163,37 181,42:

$$\bar{x} = 176.4444 \,; \bar{y} = 42.44444$$

$$b_{min} = 0.2811502 \,; a_{min} = -7.16294$$

$$Schuhgröße = -7.16294 + 0.2811502 * Körpergröße$$

oberer Teil der Formel: $\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$ Kovarianz

Dieser Wert wird größer, wenn x und y in die gleiche Richtung variieren

unterer Teil der Formel: $\sum_{i=1}^{n} (x_i - \bar{x})^2$ Varianz von X

normiert die gemeinsame Varianz auf die Varianz von x

Ergebnis: Wie variiert y in Verhältnis zu x im Durchschnitt

Regression: least-squares Methode (Methode der kleinsten Quadrate) [3]

Formel für die least-square Methode

In R:

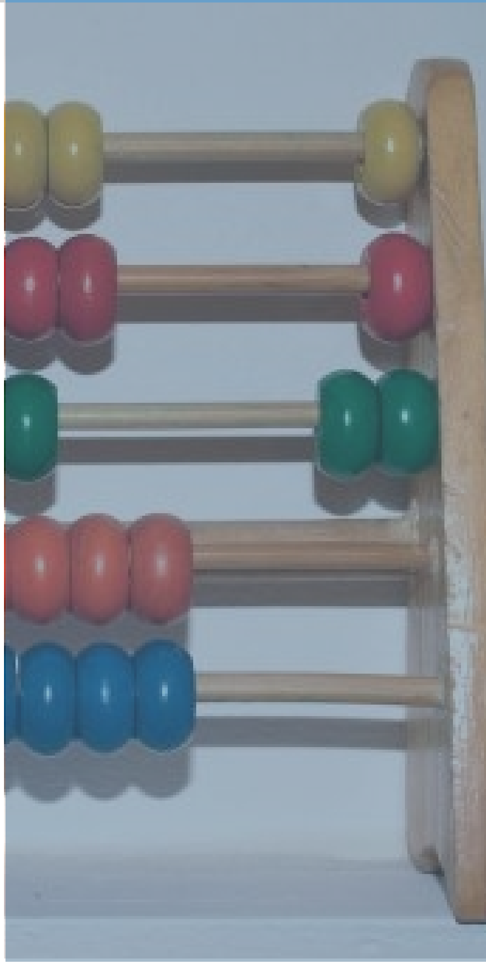
```
> b.min<-sum((kursdaten$Gr-
mean(kursdaten$Gr))*(kursdaten$Schuhgroesse-
mean(kursdaten$Schuhgroesse))) / sum((kursdaten$Gr-
mean(kursdaten$Gr))^2)
> a.min<-mean(kursdaten$Schuhgroesse)-b.min*mean(kursdaten$Gr)
> a.min
[1] -7.16294
> b.min
[1] 0.2811502
>
```

Oder kürzer:

```
> lm(kursdaten$Schuhgroesse ~ kursdaten$Gr)

Call:
lm(formula = kursdaten$Schuhgroesse ~ kursdaten$Gr)

Coefficients:
 (Intercept)  kursdaten$Gr
     -7.1629        0.2812
```

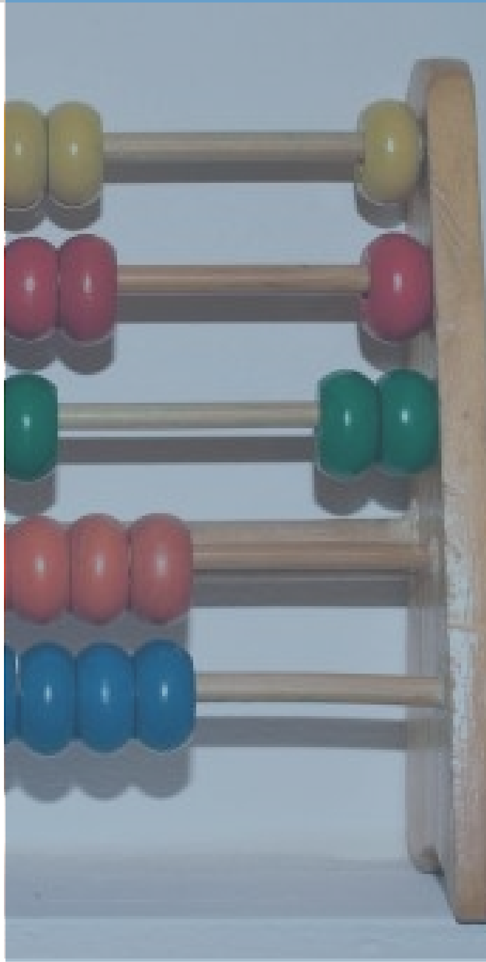


Regression: least-squares Methode Aufgabe

Regression zwischen Anzahl der Mahlsteine und Anzahl von Getreidekörnern (Beispiel nach Shennan)

Gegeben ist die Anzahl von Getreidekörnern und Mahlsteinen in verschiedenen neolithischen Siedlungen. Plotten Sie die Beziehung und geben sie die beschreibene Regressionsgleichung an.

Datei: cereal_processing.csv



Regression: least-squares Methode Lösung

Regression zwischen Anzahl der Mahlsteine und Anzahl von Getreidekörnern (Beispiel nach Shennan)

Gegeben ist die Anzahl von Getreidekörnern und Mahlsteinen in verschiedenen neolithischen Siedlungen. Plotten Sie die Beziehung und geben sie die beschreibene Regressionsgleichung an.

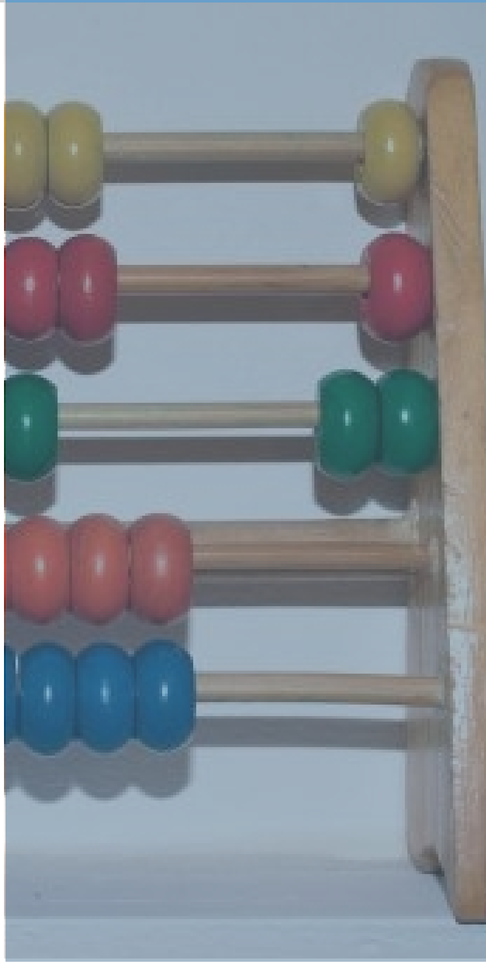
Datei: cereal_processing.csv

```
> data<-read.csv2("cereal_processing.csv",row.names=1)
> plot(data$groundstones,data$cereals)
> lm(data
data          data=         data.class    data.entry    dataentry
data.frame    data.matrix   datasets::
> lm(data$cereals ~ data$groundstones)

Call:
lm(formula = data$cereals ~ data$groundstones)

Coefficients:
      (Intercept)  data$groundstones
           122.67              39.86
```

Getreidekörner = 122.67 + 39.86 * Mahlsteine



Korrelation: Korrelationskoeffizient [1]

Wie gut passt meine Regressionsgleichung zu den Daten

Die Regression ist nur eine optimale Annäherung, deren Güte davon abhängt, wie gut die unabhängige Variable die abhängige determiniert.

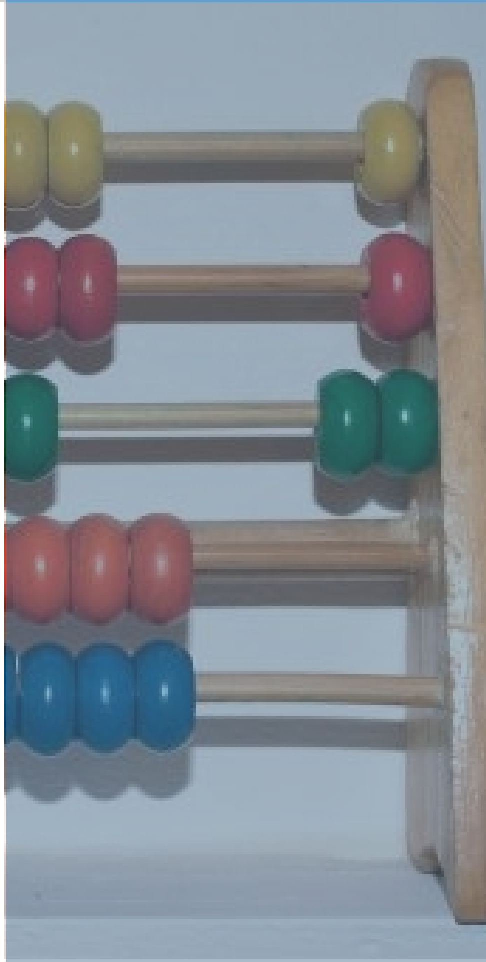
```
> abline(lm(data$cereals ~ data$groundstones))
```

In der Realität weichen die Daten meist von der Ideallinie ab.

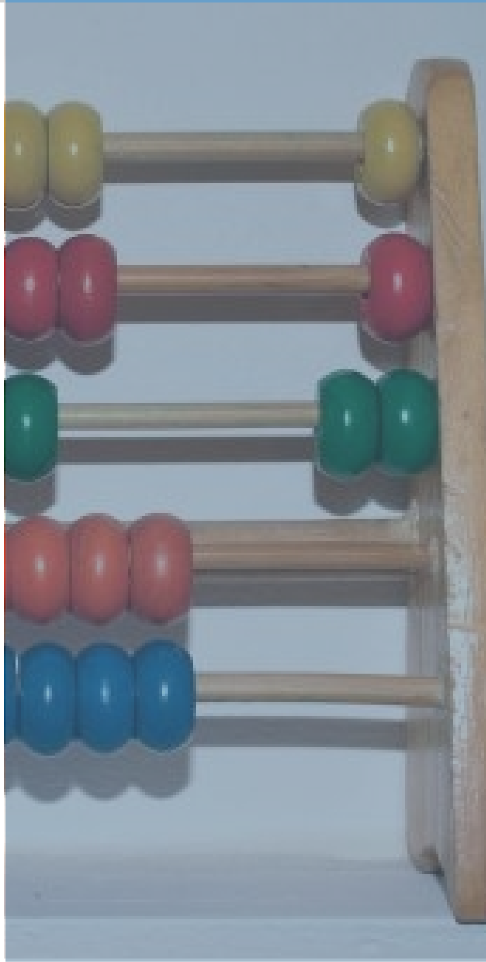
Wie stark ist also die Korrelation?

Korrelationskoeffizient:
Maß dafür wie sehr die Daten sich um die Regressionsgerade verteilen, Maß dafür wie stark die Variablen *kovariieren* in Bezug auf ihre eigene Varianz

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



Korrelation: Korrelationskoeffizient [2]



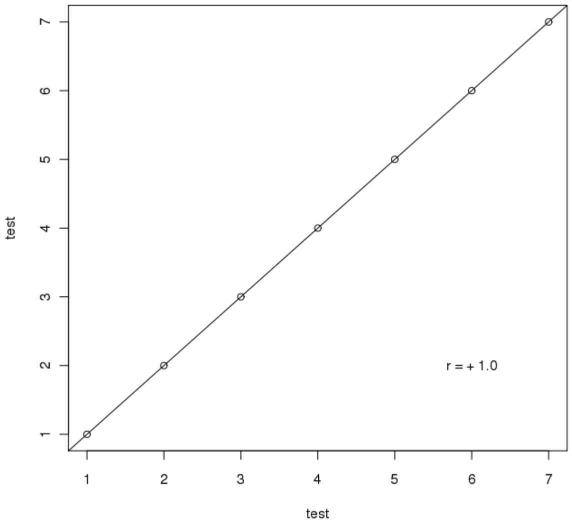
Formel

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

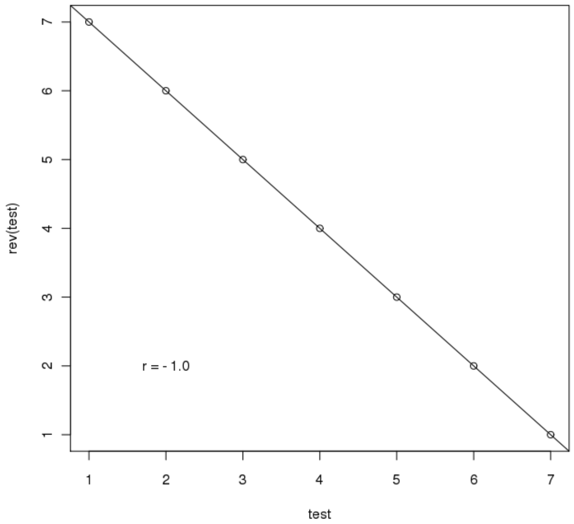
oberer Teil der Formel: $\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$ Kovarianz

unterer Teil der Formel: $\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}$ standardisiert die Kovarianz auf beide Varianzen

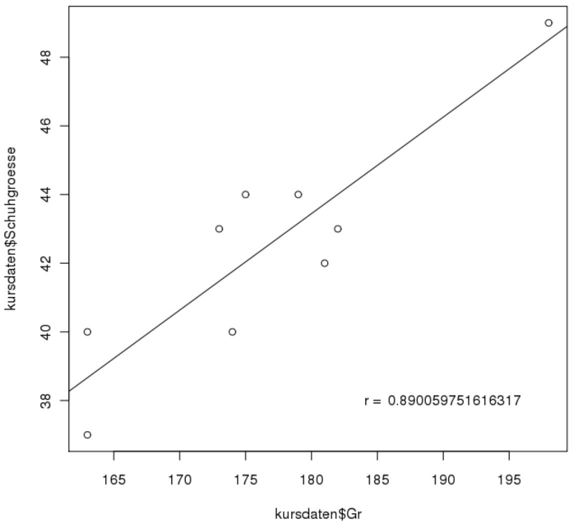
wenn die gemeinsame Varianz größer ist als die unabhängigen Varianzen steig r
wenn die gemeinsame Varianz kleiner ist als die unabhängigen Varianzen sinkt r
wenn alle Werte auf einer Linie liegen wird |r| 1
wenn x steigt und damit y steigt wird der Wert positiv
wenn x steigt und damit y sinkt wird der Wert negativ



Perfekte Positive Korrelation



Perfekte negative Korrelation



Beispieldaten
Schuhgröße gg. Körpergröße

Korrelation: Korrelationskoeffizient [3]

In R:

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

oberer Teil der Formel: $\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$ Kovarianz

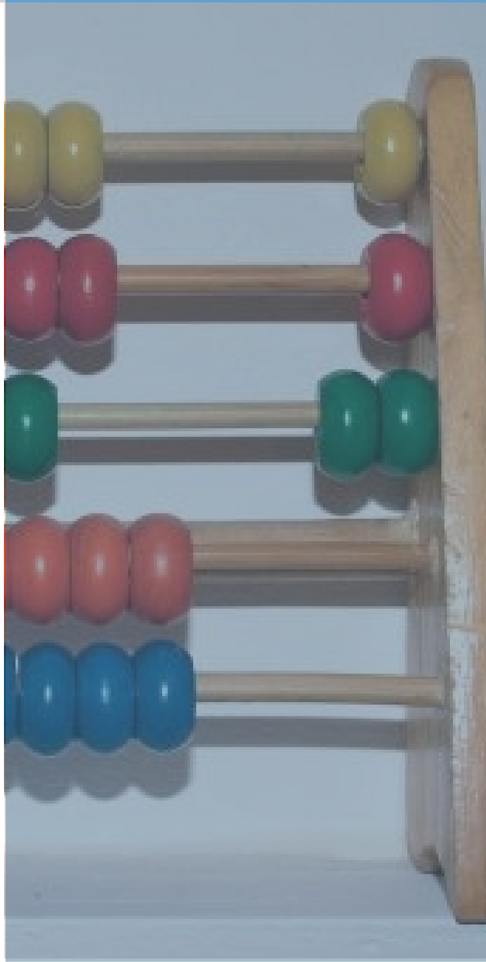
unterer Teil der Formel: $\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}$ standardisiert die Kovarianz auf beide Varianzen

```
cov(kursdaten$Gr,kursdaten$Schuhgroesse)/sqrt(var(kursdaten$Gr)*var(
kursdaten$Schuhgroesse))
[1] 0.8900598
```

Kovarianz (cov) durch Wurzel aus (Varianz Gr * Varianz Schuhgröße)

Oder einfacher:

```
> cor (kursdaten$Gr,kursdaten$Schuhgroesse)
[1] 0.8900598
```



Korrelation: Bestimmtheitsmaß (Determinationskoeffizient)

Gibt an, wieviel der Variation der abhängigen Variable durch die Variation der unabhängigen Variable erklärt wird

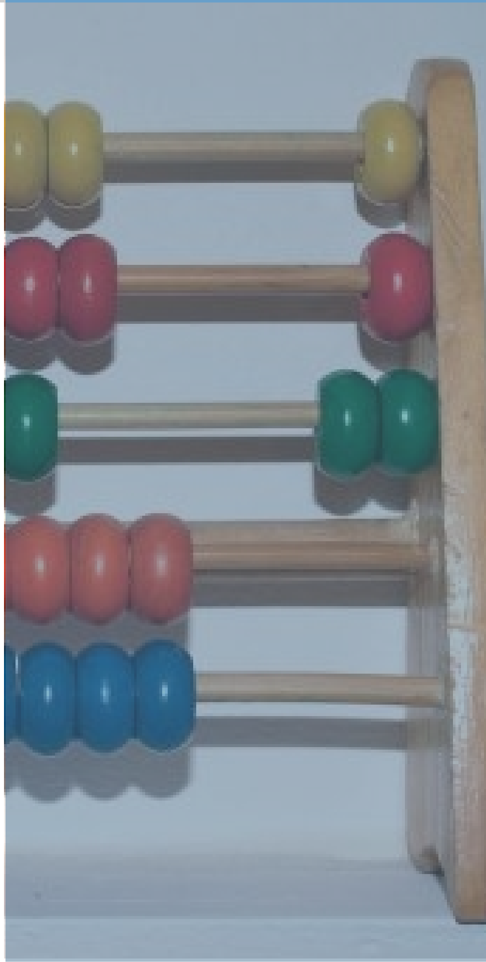
Beispiel: zu wieviel Prozent ist die Schuhgröße durch die Körpergröße erklärt?

Determinationskoeffizent $r^2 = r$ ^2 ;-)

Unser Beispiel: $r = 0.8900598$, $r^2 = 0.7922064$

79,22 % der Variation in der Schuhgröße wird durch die Körpergröße erklärt!

Achtung: “erklärt“ heißt nicht zwingend kausaler Zusammenhang!



Test auf Korrelation

Es korreliert, aber korreliert es auch signifikant?

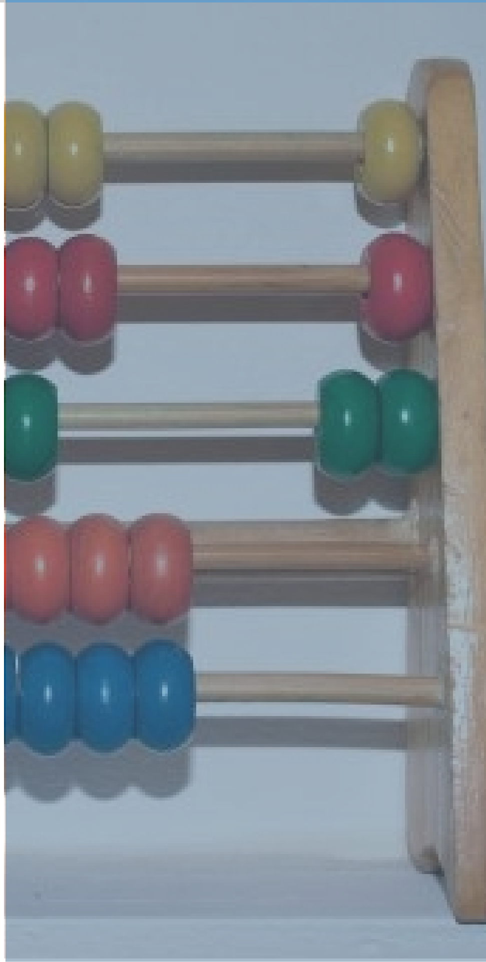
Test gegen eine normalverteilte Fehlerverteilung mit Pearsons Korrelationskoeffizent (der „normale“ Korrelationskoeffizient)

Die Variablen sollten normalverteilt sein (überprüfen mit ks.test oder shapiro.test)

```
> shapiro.test(kursdaten$Gr)
…
W = 0.9189, p-value = 0.3834
> shapiro.test(kursdaten$Schuhgroesse)
…
W = 0.9481, p-value = 0.6694
> cor.test(kursdaten$Gr,kursdaten$Schuhgroesse)

     Pearson's product-moment correlation

data:  kursdaten$Gr and kursdaten$Schuhgroesse
t = 5.166, df = 7, p-value = 0.001301
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5525617 0.9767919
sample estimates:
      cor
0.8900598
```

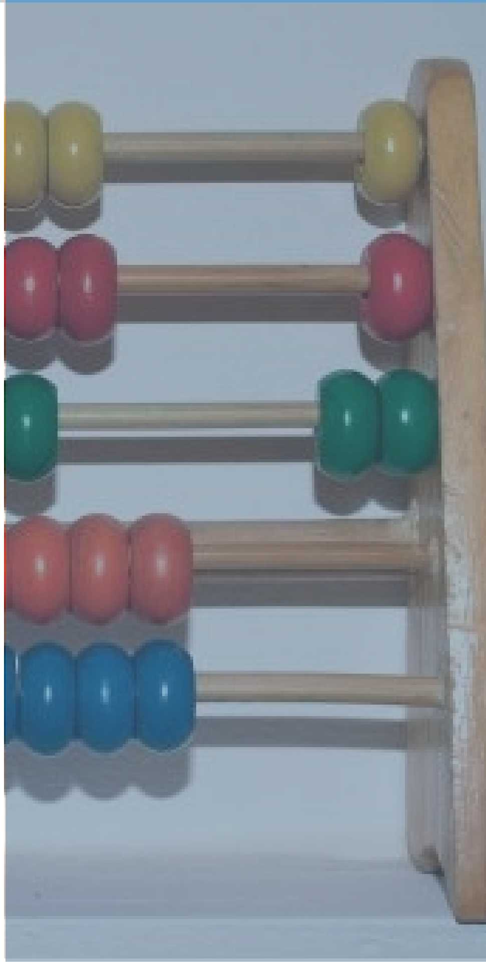


Korrelation: Aufgabe

Korrelation zwischen Anzahl der Mahlsteine und Anzahl von Getreidekörnern (Beispiel nach Shennan)

Gegeben ist die Anzahl von Getreidekörnern und Mahlsteinen in verschiedenen neolithischen Siedlungen. Geben Sie an, wie stark die Variablen miteinander korrelieren, wieviel der Variation der Mahlsteine durch die Getreidekörner erklärt wird und ob die Korrelation signifikant ist!

Datei: cereal_processing.csv



Korrelation: Lösung

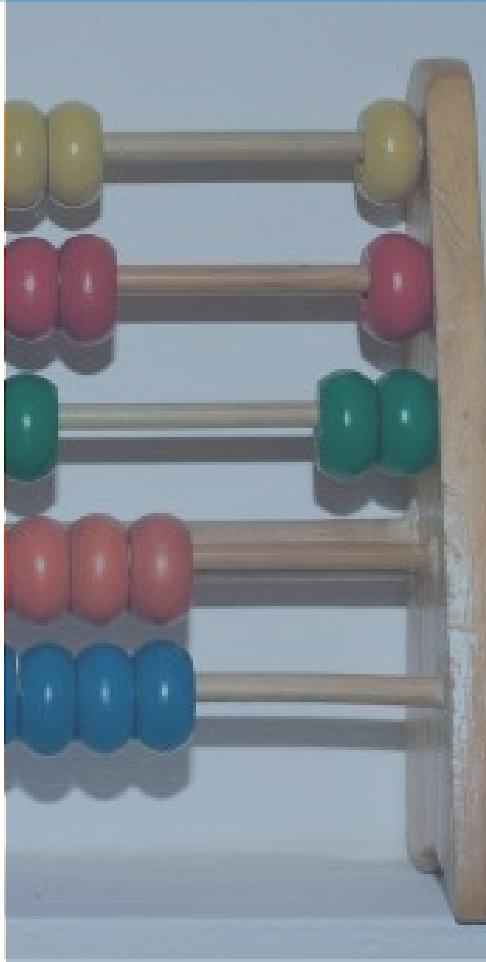
Korrelation zwischen Anzahl der Mahlsteine und Anzahl von Getreidekörnern (Beispiel nach Shennan)

Gegeben ist die Anzahl von Getreidekörnern und Mahlsteinen in verschiedenen neolithischen Siedlungen. Geben Sie an, wie stark die Variablen miteinander korrelieren und wieviel der Variation der Mahlsteine durch die Getreidekörner erklärt wird!

Datei: cereal_processing.csv

```
> cor(data$groundstones,data$cereals)
[1] 0.790843
> cor(data$groundstones,data$cereals)^2
[1] 0.6254327
> shapiro.test(data$cereals)
> shapiro.test(data$groundstones)
> cor.test(data$groundstones,data$cereals)
…
t = 4.2857, df = 11, p-value = 0.001286
```

Die Mahlsteine korrelieren signifikant positiv mit den Getreidekörnern. Die Anzahl der Mahlsteine wird zu 62,54 % durch die Anzahl der Getreidekörner erklärt.



Korrelation ordinal skalierter Variablen

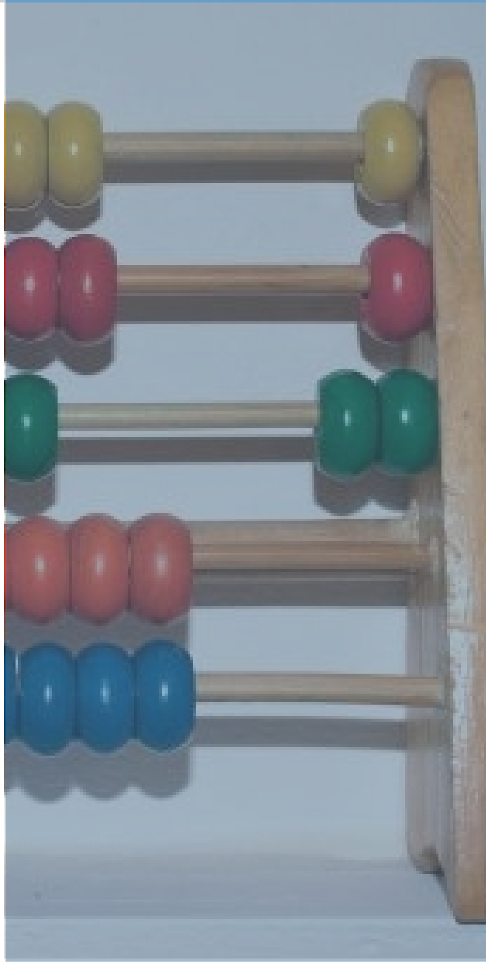
Wenn wir, wie so oft, keine Messdaten haben

Maße für Korrelation ordinal skalierter Daten (Rankkorrelation):

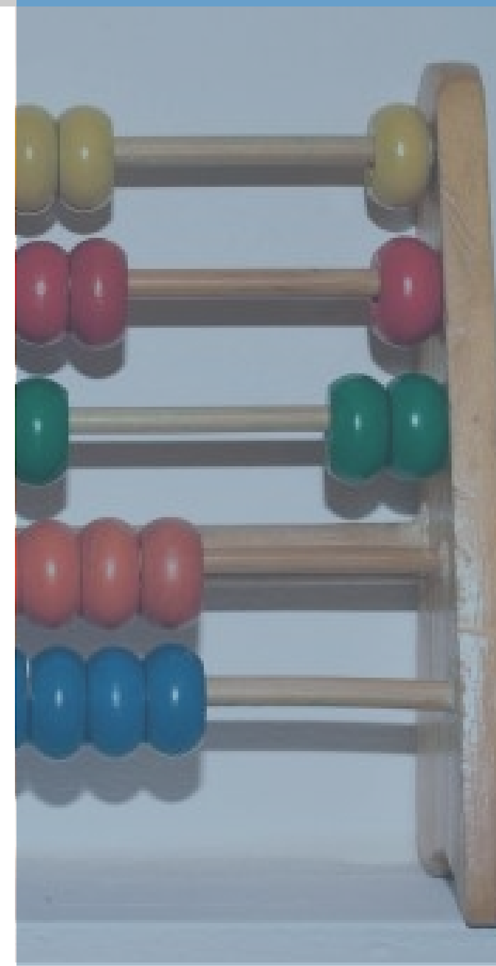
Kendall's τ
Spearmans rho

Beispiel nach Shennan: Größe der Siedlung und Qualität des Bodens

Größe	Bodenqualität			Gesamt
	Hervorragend	Durchschnitt	Arm	
Groß	15	7	2	24
Mittel	6	11	4	21
Klein	7	7	8	22
Gesamt	28	25	14	67



Kendall's τ [1]



Berechnung über die Ränge

Voraussetzungen: Zwei mind. ordinal skalierte Variablen einer Stichprobe

Idee: Bei einer perfekten Korrelation sind alle großen Siedlungen auf den guten Böden, alle mittelgroßen auf den mittleren und alle kleinen auf den schlechten.

Berechnet wird mit mögliche Paarungen von Werten, deren Ränge zueinander werden verglichen.

Wenn für eine Paarung sowohl x als auch y-Wert kleiner ist als bei bei dem Vergleichspaar, dann ergibt sich eine konkordierende Paarung (Bei beiden stimmt die Rangfolge überein).

Wenn für eine Paarung der x-Wert größer, aber der y-Wert kleiner ist, dann handelt es sich um ein diskonkordierendes Paar.

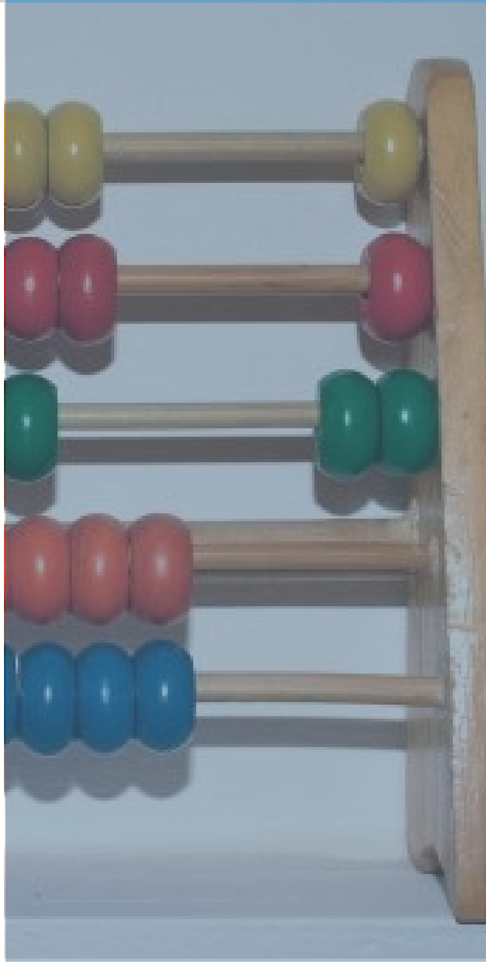
Kendall's τ_c [1]

Berechnung konkordierender Ränge

Größe	Bodenqualität			Gesamt
	Hervorragend	Durchschnitt	Arm	
Groß	15(a)	7 (d)	2 (g)	24
Mittel	6 (b)	11 (e)	4 (h)	21
Klein	7 (c)	7 (f)	8 (i)	22
Gesamt	28	25	14	67

Alle Siedlungen in Zelle a können mit allen Siedlungen in e,f,h,i kombiniert werden, so dass sowohl Bodenqualität als auch Siedlungsgröße in a größer sind als in e,f,h,i.

Paarungen: a*(e+f+h+i)= 15*(11+7+4+8)=450



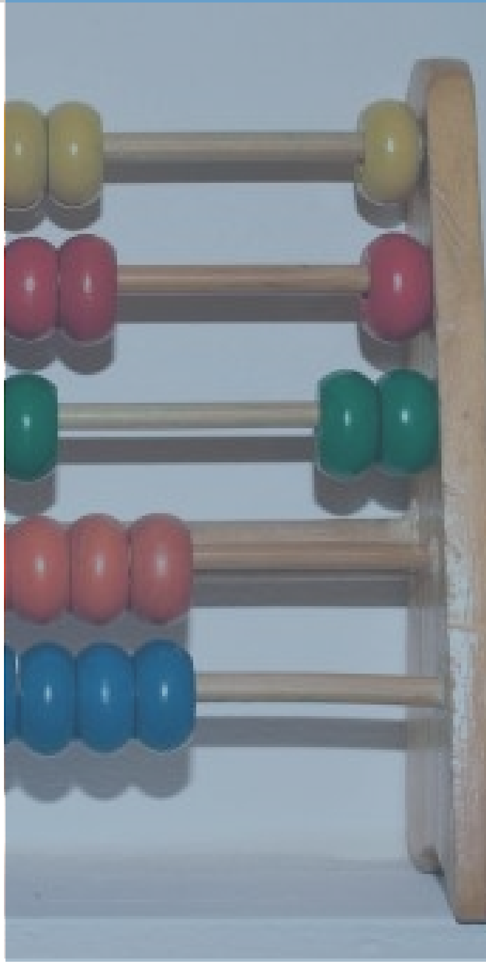
Kendall's τ_c [2]

Berechnung konkordierender Ränge

Größe	Bodenqualität			Gesamt
	Hervorragend	Durchschnitt	Arm	
Groß	15(a)	7 (d)	2 (g)	24
Mittel	6 (b)	11 (e)	4 (h)	21
Klein	7 (c)	7 (f)	8 (i)	22
Gesamt	28	25	14	67

Alle Siedlungen in Zelle b können mit allen Siedlungen in f,i kombiniert werden, so dass sowohl Bodenqualität als auch Siedlungsgröße in a größer sind als in f,i.

Paarungen: b*(f+i)= 6*(7+8)=90



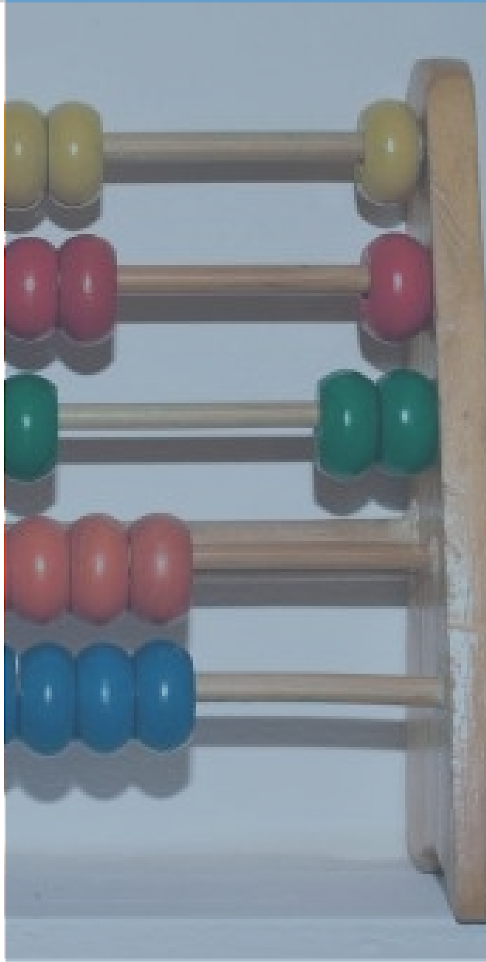
Kendall's τ_c [3]

Berechnung konkordierender Ränge

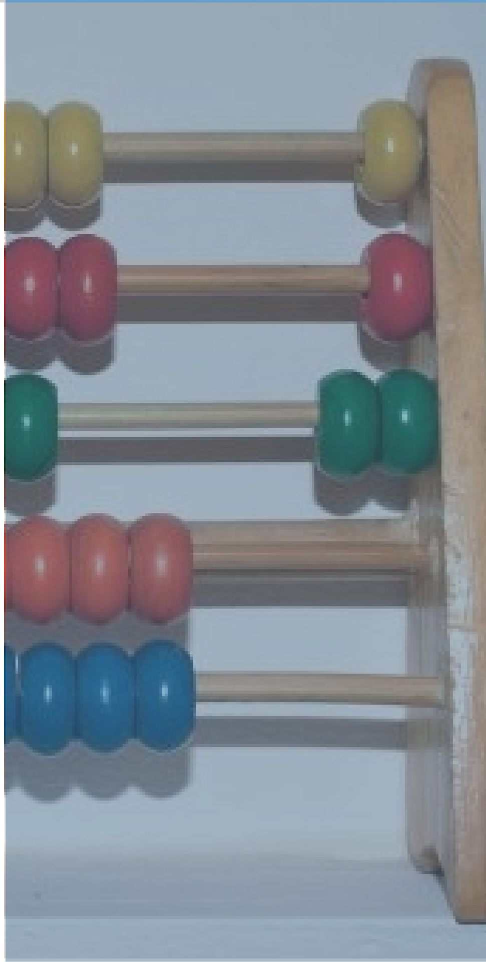
Größe	Bodenqualität			Gesamt
	Hervorragend	Durchschnitt	Arm	
Groß	15(a)	7 (d)	2 (g)	24
Mittel	6 (b)	11 (e)	4 (h)	21
Klein	7 (c)	7 (f)	8 (i)	22
Gesamt	28	25	14	67

Alle Siedlungen in Zelle d können mit allen Siedlungen in h,i kombiniert werden, so dass sowohl Bodenqualität als auch Siedlungsgröße in a größer sind als in h,i.

Paarungen: d*(h+i)= 7*(4+8)=84



Kendall's τ_c [4]



Berechnung konkordierender Ränge

Größe	Bodenqualität			Gesamt
	Hervorragend	Durchschnitt	Arm	
Groß	15(a)	7 (d)	2 (g)	24
Mittel	6 (b)	11 (e)	4 (h)	21
Klein	7 (c)	7 (f)	8 (i)	22
Gesamt	28	25	14	67

Alle Siedlungen in Zelle e können mit allen Siedlungen in i kombiniert werden, so dass sowohl Bodenqualität als auch Siedlungsgröße in a größer sind als in i.

Paarungen: e*i= 11*8=88

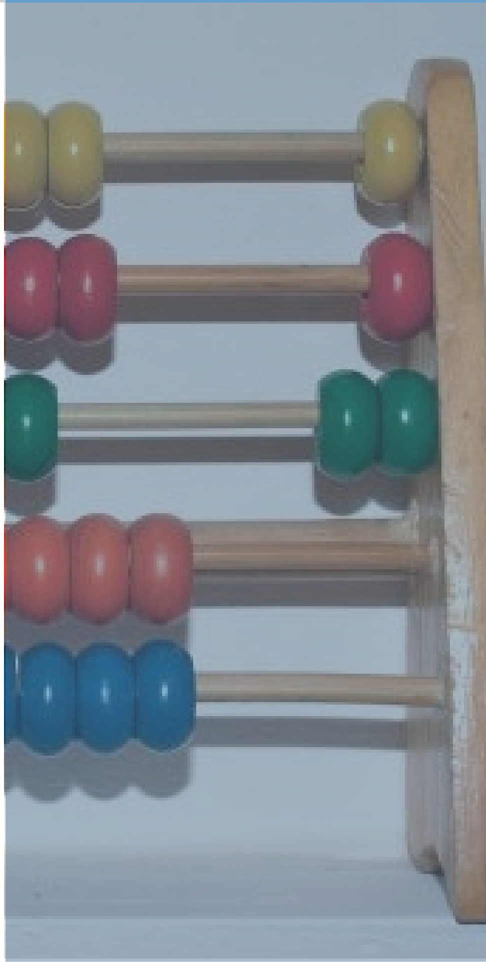
Kendall's τ_c [5]

Berechnung konkordierender Ränge

Größe	Bodenqualität			Gesamt
	Hervorragend	Durchschnitt	Arm	
Groß	15(a)	7 (d)	2 (g)	24
Mittel	6 (b)	11 (e)	4 (h)	21
Klein	7 (c)	7 (f)	8 (i)	22
Gesamt	28	25	14	67

Die Anzahl der Paarungen mit konkordierenden Rängen ist also die Summe der einzelnen möglichen Paarungen.

Paarungen: C=450+90+84+88=712



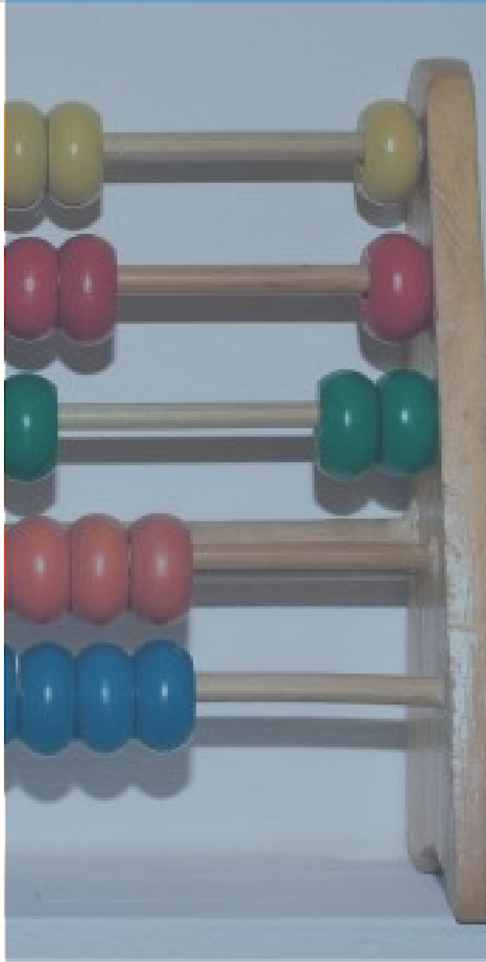
Kendall's τ_c [6]

Berechnung diskonkordierender Ränge

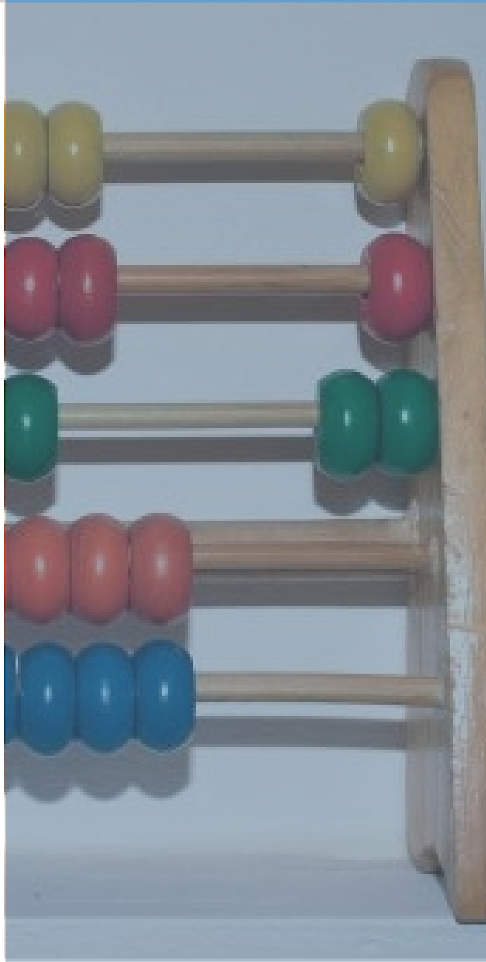
Größe	Bodenqualität			Gesamt
	Hervorragend	Durchschnitt	Arm	
Groß	15(a)	7 (d)	2 (g)	24
Mittel	6 (b)	11 (e)	4 (h)	21
Klein	7 (c)	7 (f)	8 (i)	22
Gesamt	28	25	14	67

Alle Siedlungen in Zelle g können mit allen Siedlungen in b,c,e,f kombiniert werden, so dass Bodenqualität schlechter, Siedlungsgröße aber größer ist als in b,c,e,f.

Paarungen: g*(b+c+e+f)=2*(6+11+7+7)=62



Kendall's τ_c [7]



Berechnung diskonkordierender Ränge

Größe	Bodenqualität			Gesamt
	Hervorragend	Durchschnitt	Arm	
Groß	15(a)	7 (d)	2 (g)	24
Mittel	6 (b)	11 (e)	4 (h)	21
Klein	7 (c)	7 (f)	8 (i)	22
Gesamt	28	25	14	67

Alle Siedlungen in Zelle h können mit allen Siedlungen in c,f kombiniert werden, so dass Bodenqualität schlechter, Siedlungsgröße aber größer ist als in c,f.

Paarungen: h*(c+f)=4*(7+7)=56

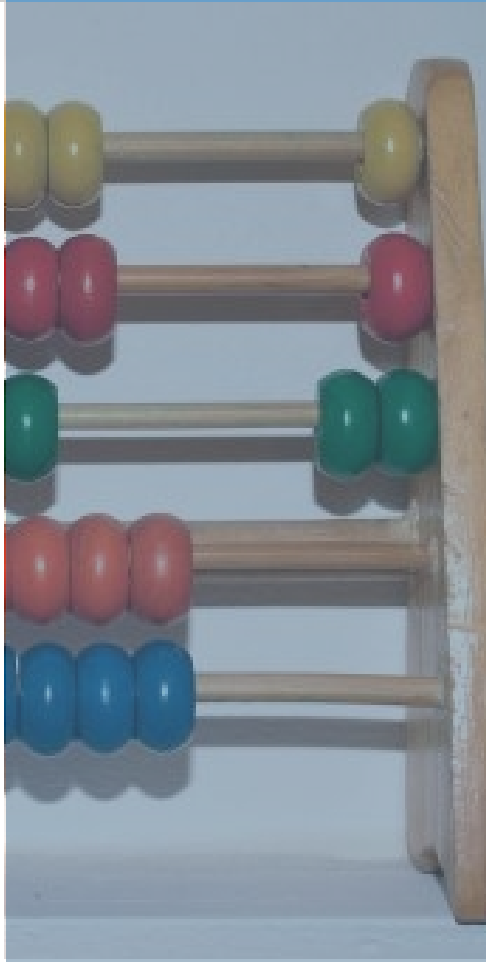
Kendall's τ_c [8]

Berechnung diskonkordierender Ränge

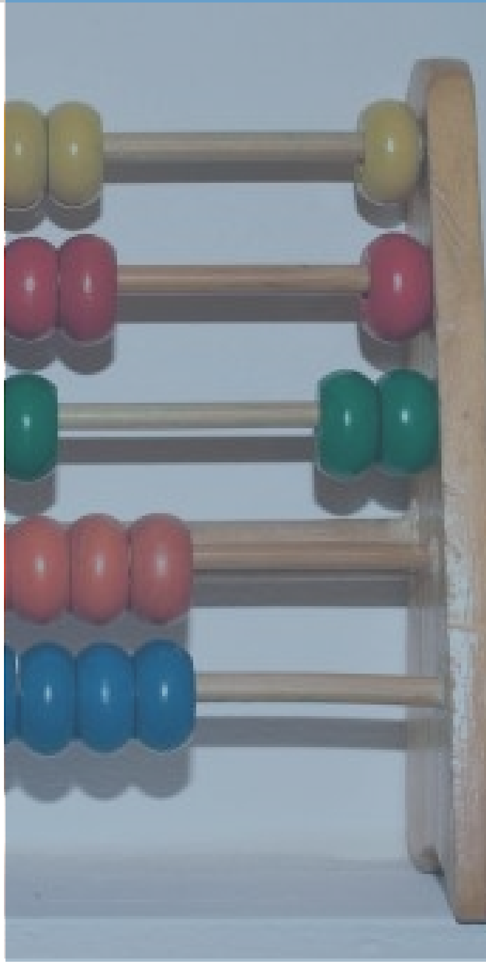
Größe	Bodenqualität			Gesamt
	Hervorragend	Durchschnitt	Arm	
Groß	15(a)	7 (d)	2 (g)	24
Mittel	6 (b)	11 (e)	4 (h)	21
Klein	7 (c)	7 (f)	8 (i)	22
Gesamt	28	25	14	67

Alle Siedlungen in Zelle d können mit allen Siedlungen in b,c kombiniert werden, so dass Bodenqualität schlechter, Siedlungsgröße aber größer ist als in b,c.

Paarungen: d*(b+c)=7*(6+7)=91



Kendall's τ_c [9]



Berechnung diskonkordierender Ränge

Größe	Bodenqualität			Gesamt
	Hervorragend	Durchschnitt	Arm	
Groß	15(a)	7 (d)	2 (g)	24
Mittel	6 (b)	11 (e)	4 (h)	21
Klein	7 (c)	7 (f)	8 (i)	22
Gesamt	28	25	14	67

Alle Siedlungen in Zelle e können mit allen Siedlungen in c kombiniert werden, so dass Bodenqualität schlechter, Siedlungsgröße aber größer ist als in c.

Paarungen: e*c=11*7=77

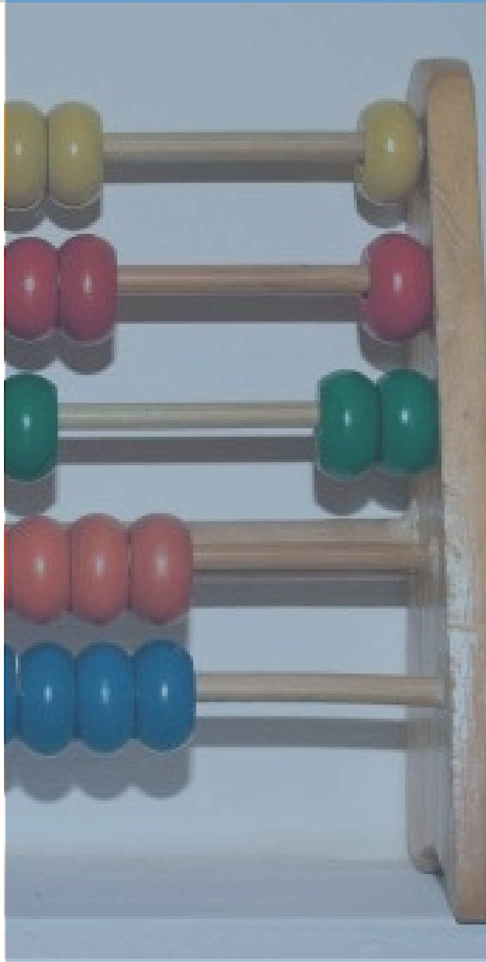
Kendall's τ_c [1]

Berechnung diskonkordierender Ränge

Größe	Bodenqualität			Gesamt
	Hervorragend	Durchschnitt	Arm	
Groß	15(a)	7 (d)	2 (g)	24
Mittel	6 (b)	11 (e)	4 (h)	21
Klein	7 (c)	7 (f)	8 (i)	22
Gesamt	28	25	14	67

Die Anzahl der Paarungen mit diskonkordierenden Rängen ist also die Summe der einzelnen möglichen Paarungen.

Paarungen: D=62+56+91+77=286



Kendall's τ_c [10]

Formel und Berechnung

$$\tau_c = \frac{C-D}{\frac{1}{2} n^2 \frac{m-1}{m}} \quad mit\ m = min(n_{zeilen}, n_{spalten})$$

$$n=67\ ; C=712\ ; D=286\ ; m=3$$

$$\tau_c = \frac{712-286}{\frac{1}{2} 64^2 \frac{3-1}{3}}$$

$$\tau_c = \frac{426}{1496{,}3}$$

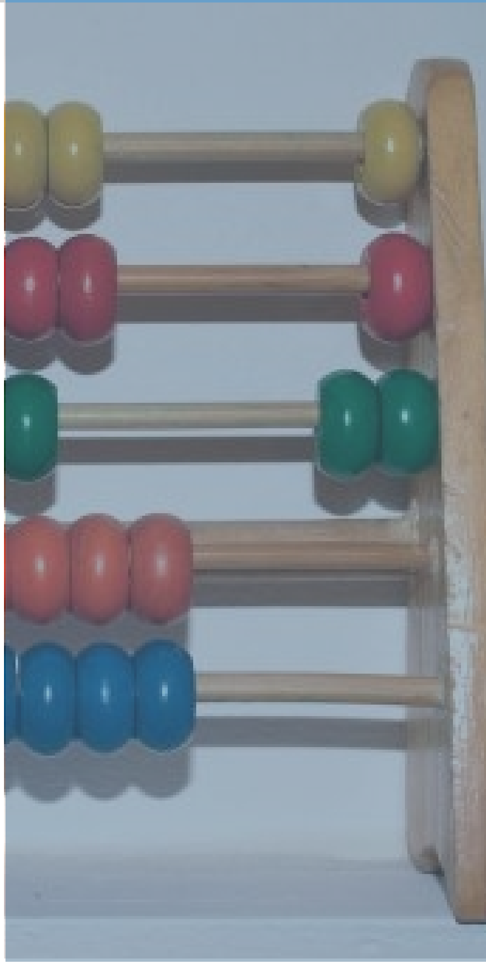
$$\tau_c = 0.285$$

In R:
Achtung: es gibt keine Berechnung für Kendall's tau c, nur für Kendall's tau b in R. Daher müssen die Daten pur, nicht als Kontingenztabelle vorliegen.

```
> soilsites<-read.csv2("soilsites.csv",row.names=1)
> cor.test(soilsites$groesse,soilsites$bodenguete,method="kendall")

     Kendall's rank correlation tau

data:  soilsites$groesse and soilsites$bodenguete
z = 2.6372, p-value = 0.008359
alternative hypothesis: true tau is not equal to 0
sample estimates:
      tau
0.2902363
```



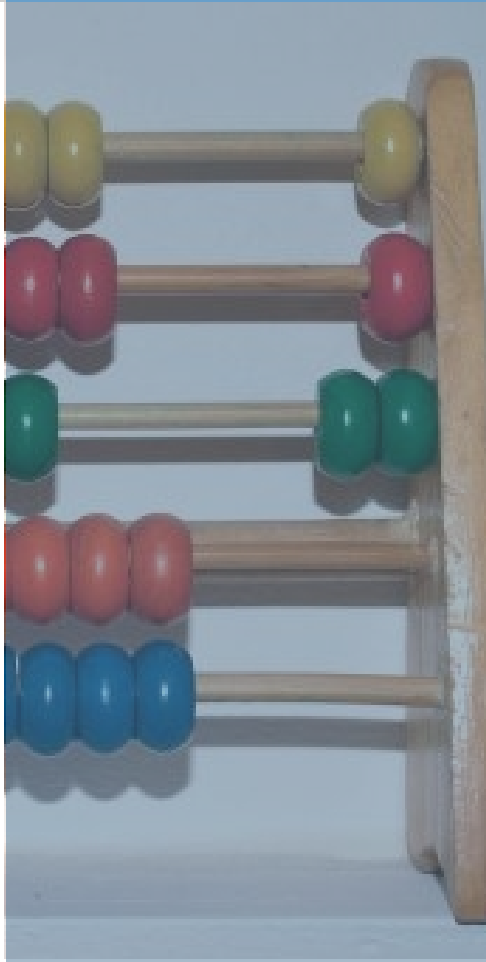
Zu Bedenken:

Korrelation ist nicht automatisch kausaler Zusammenhang!

Beispiel: Das allseits bekannte Klapperstorchbeispiel

Die Rückgang der Störche korreliert mit dem Rückgang der Geburten in Deutschland... kausaler Zusammenhang?

Oft sind es versteckte komplexe dritte Variablen, die zwei korrelierende Größen beeinflussen, z.B. die Veränderungen in der modernen Gesellschaft, die sowohl Rückgang der Störche wie auch der Geburten beeinflussen.



Nächste Sitzung 27. Januar 2011:
Clusteranalyse

Bitte lesen Sie Shennan 8 „Relationships...“ und besonders 9 „When the Regression Doesn't Fit“.