UNIVERSITÄT BERN

Statistical methods for archaeological data analysis I: Basic methods

08 - Regression and Correlation

Martin Hinz

Institut für Archäologische Wissenschaften, Universität Bern

28.04.21

Loading data for the following steps

Read the Data on Muensingen Fibulae

muensingen_fib.csv

```
muensingen <- read.csv2("muensingen_fib.csv")
head(muensingen)
```

```
##    X Grave Mno FL BH BFA FA CD BRA ED FEL  C   BW  BT FEW Coils Length
## 1  1   121 348 28 17   1 10 10   2  8   6 20  2.5 2.6 2.2     4     53
## 2  2   130 545 29 15   3  8  6   3  6  10 17 11.7 3.9 6.4     6     47
## 3  3   130 549 22 15   3  8  7   3 13   1 17  5.0 4.6 2.5    10     47
## 4  8   157  85 23 13   3  8  6   2 10   7 15  5.2 2.7 5.4    12     41
## 5 11   181 212 94 15   7 10 12   5 11  31 50  4.3 4.3  NA     6    128
## 6 12   193 611 68 18   7  9  9   7  3  50 18  9.3 6.5  NA     4    110
##   fibula_scheme
## 1             B
## 2             B
## 3             B
## 4             B
## 5             C
## 6             C
```

Scatterplot

For 2 variables

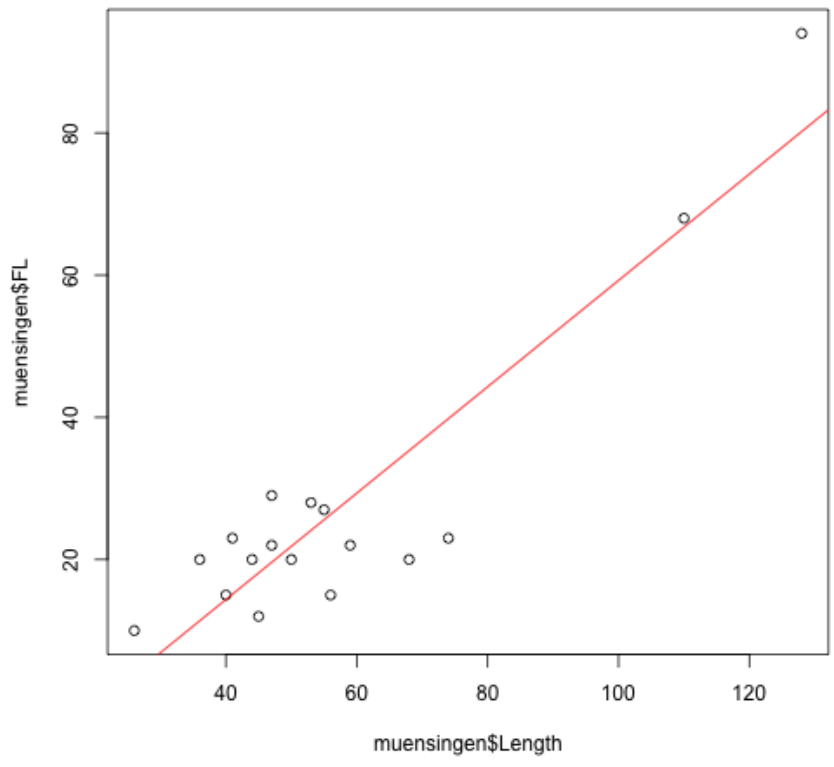
Used to display a variable in relation to another one.

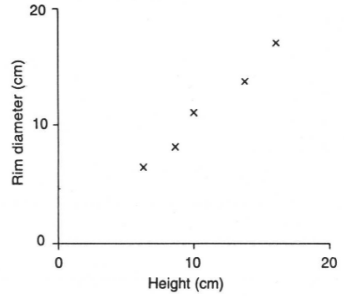
```
plot(muensingen$Length, muensingen$FL)
abline(
  lm(muensingen$FL~muensingen$Length),
  col="red")
```

Visible: If one variable increases in size, the other variable increases as well.

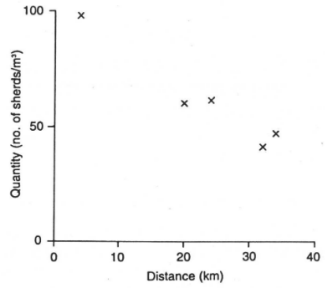
Other visible properties:

- Direction of the relationship (greater-> greater vs. greater -> smaller)
- Linearity of the relation (monotonous, not monotonous)
- Strength of the relationship (points near vs. far from an imaginary line)

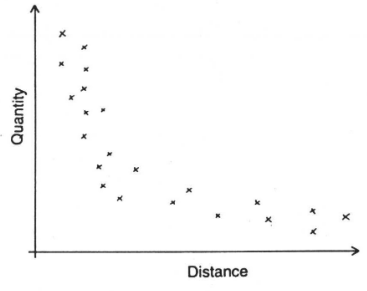




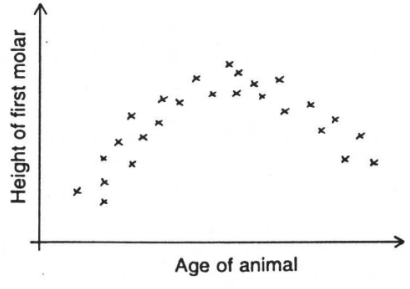
positive regression



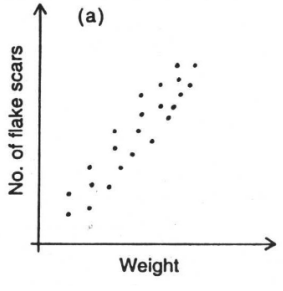
negative regression



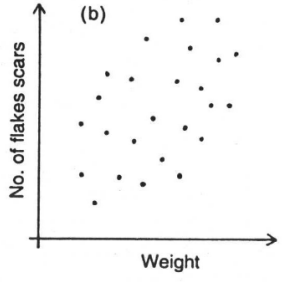
nonlinear
monotone regression



nonlinear
nonmonotone regression



strong correlation



weak correlation

Direction and linearity of relationships

direction

Indicates whether a variable increases (positive) or decreases (negative) with the other variable.

Variables: possible cause (independent variable) and effect of interest (dependent variable)

linearity

There are linear and non-linear regressions.

Non-linear regressions, possible causes:

Combination of different (linear?) influences: multiple regression analysis

Influence factor has no linear effect: nonlinear model (square or higher polynomial, threshold systems etc.)

Regression: Equation

What we still know from school lessons... The formula for a linear equation consists of a slope (b) and an intercept (displacement constant a)

$$y = a + bx$$

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

$$a = y_1 - b * x_1$$

Example: {1,3}, {2,5}, {3,7} ...

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

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

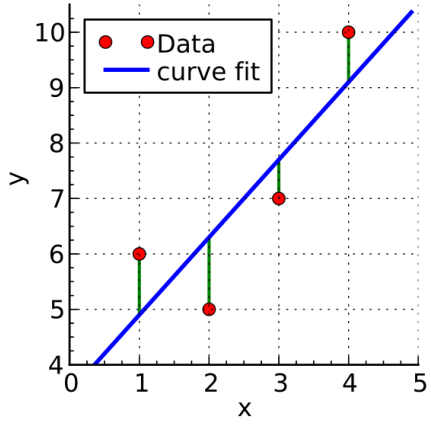
$$y = 1 + 2 * x$$

But: this only works with perfect correlation, what with deviating (statistical) values?

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

Estimation of the optimal approximation with the least-square method

For values that do not correspond exactly to a straight line, an optimal approximation must be found.



https://commons.wikimedia.org/wiki/File:Linear_least_squares_example2.svg

The absolute distance between the real y-value and the estimated y-value should be as small as possible, it applies:

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

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

slope

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

upper part of the formula:
$\sum_{i=1}^{n} (x_i - \bar{x}) * (y_i - \bar{y})$ **covariance**

This value increases when x and y vary in the same direction.

lower part of the formula: $\sum_{i=1}^{n} (x_i - \bar{x})^2$
variance of X

normalizes the common variance to the variance of x

Result: How does y vary in relation to x on average?

intercept

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

given the slope, what is the displacement (intercept with the y-axis) in respect to the means of both variables

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

Example

```
head(muensingen[,c("FL", "Length")])
```

```
##   FL Length
## 1 28     53
## 2 29     47
## 3 22     47
## 4 23     41
## 5 94    128
## 6 68    110
```

```
colMeans(head(muensingen[,c("FL", "Length")]))
```

```
##     FL Length
##     44     71
```

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

FL	Length	FL - mean(FL)	Length - mean(Length)	covariance	var(L)
28	53	-16	-18	288	324
29	47	-15	-24	360	576
22	47	-22	-24	528	576
23	41	-21	-30	630	900
94	128	50	57	2850	3249
68	110	24	39	936	1521
sum				5592	7146

$$b_{min} = \frac{5592}{7146}$$

$$b_{min} = 0.7825357$$

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

$$a_{min} = 44 - 0.7825357 * 71$$

$$a_{min} = -11.5600336$$

Regression: least-squares Methode (Methode der kleinsten Quadrate) [4]

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

FL	Length	FL - mean(FL)	Length - mean(Length)	covariance	var(L)
28	53	-16	-18	288	324
29	47	-15	-24	360	576
22	47	-22	-24	528	576
23	41	-21	-30	630	900
94	128	50	57	2850	3249
68	110	24	39	936	1521
sum				5592	7146

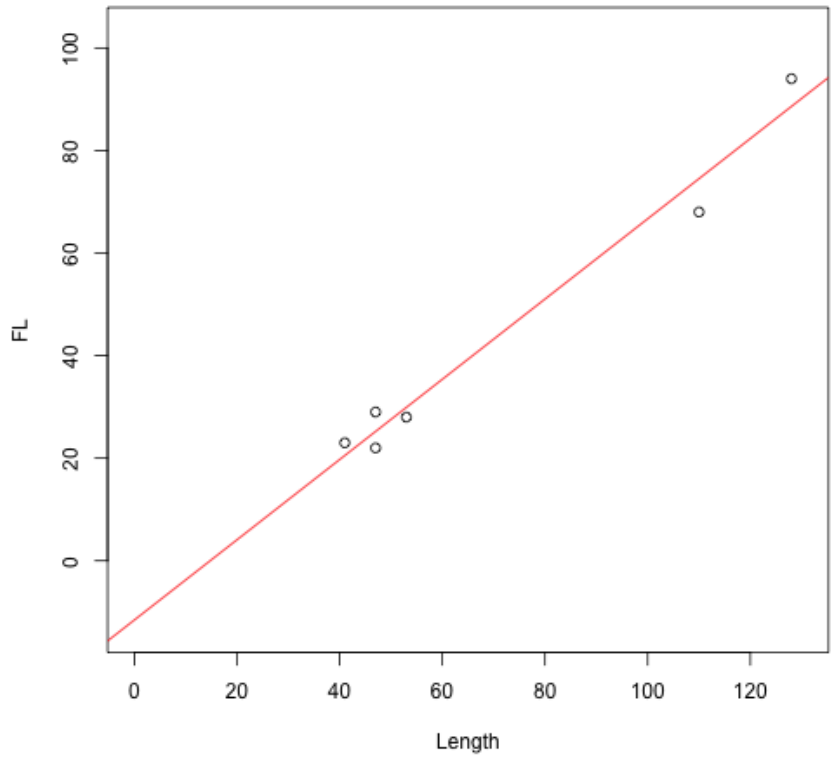
$$b_{min} = \frac{5592}{7146}$$

$$b_{min} = 0.7825357$$

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

$$a_{min} = 44 - 0.7825357 * 71$$

$$a_{min} = -11.5600336$$



Regression: least-squares Methode (Methode der kleinsten Quadrate) [5]

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

FL	Length	FL - mean(FL)	Length - mean(Length)	covariance	var(L)
28	53	-16	-18	288	324
29	47	-15	-24	360	576
22	47	-22	-24	528	576
23	41	-21	-30	630	900
94	128	50	57	2850	3249
68	110	24	39	936	1521
sum				5592	7146

$$b_{min} = \frac{5592}{7146}$$

$$b_{min} = 0.7825357$$

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

$$a_{min} = 44 - 0.7825357 * 71$$

$$a_{min} = -11.5600336$$

```
mm <- data.frame(head(muensingen))
b.min <- sum(
  (mm$FL - mean(mm$FL)) * (mm$Length - mean(mm$Length))
  ) /
  sum((mm$Length - mean(mm$Length))^2)
b.min
```

```
## [1] 0.7825357
```

```
a.min <- mean(mm$FL) - b.min * mean(mm$Length)
a.min
```

```
## [1] -11.56003
```

Or shorter:

```
lm(FL ~ Length, data=mm)
```

```
##
## Call:
## lm(formula = FL ~ Length, data = mm)
##
## Coefficients:
## (Intercept)       Length
##    -11.5600       0.7825
```

Regression: least-squares method exercise

Regression between number of millstones and number of cereal grains (Shennan example)

The number of cereal grains and millstones is given in different Neolithic settlements. Plot the relationship and specify the described regression equation.

File: cereal_processing.csv

Correlation: Correlation coefficient [1]

How well does my regression equation fit the data?

Regression is only an optimal approximation, the quality of which depends on it. depends on how well the independent variable determines the dependent one.

In reality, the data usually deviate from the ideal line.

So how strong is the correlation?

```
data<-read.csv2("cereal_processing.csv",row.names=1)
plot(data$groundstones,data$cereals)
abline(lm(data$cereals ~ data$groundstones))
```

Correlation coefficient:

Measure of how much the data is distributed around the regression line,

measure of how strongly the variables *covariate* in relation to their own variability

Correlation: Correlation coefficient [2]

Correlation coefficient:

Measure of how much the data is distributed around the regression line,

measure of how strongly the variables *covariate* in relation to their own variability

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

upper part of the formula:
$\sum_{i=1}^{n}(x_i - \bar{x}) * (y_i - \bar{y})$ covariance

lower part of the formula:
$\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 * \sum_{i=1}^{n}(y_i - \bar{y})^2}$
standardizes the covariance to both variances

```
data<-read.csv2("cereal_processing.csv",row.names=1)
plot(data$groundstones,data$cereals)
abline(lm(data$cereals ~ data$groundstones))
```

Correlation: Correlation coefficient [3]

Correlation coefficient:

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

```
data<-read.csv2("cereal_processing.csv",row.names=1)
plot(data$groundstones,data$cereals)
abline(lm(data$cereals ~ data$groundstones))
```

if the common variance is greater than the independent variances increases r

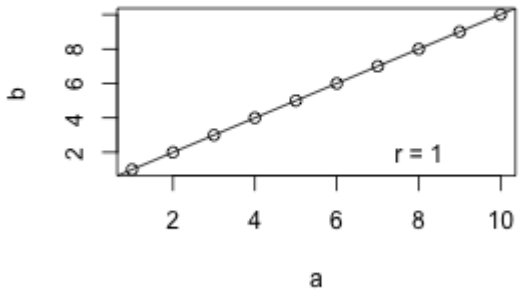
if the common variance is smaller than the independent variances r decreases

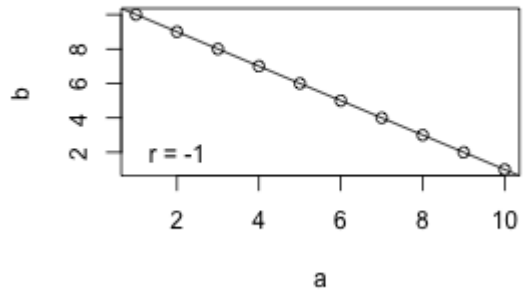
if all values lie on one line |r| = 1

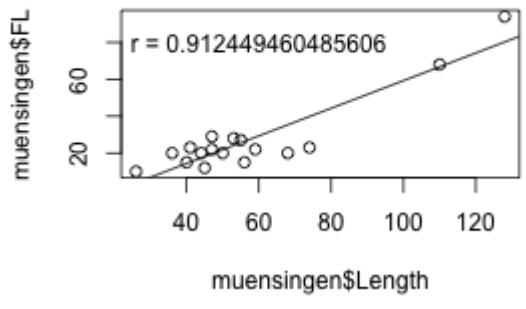
if x increases and y increases the value becomes positive

if x increases and y decreases the value becomes negative

Correlation: Correlation coefficient [4]







Correlation: Correlation coefficient [4]

in R:

Measure of how much the data is distributed around the regression line,

measure of how strongly the variables *covariate* in relation to their own variability

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

upper part of the formula:
$\sum_{i=1}^{n}(x_i - \bar{x}) * (y_i - \bar{y})$ covariance

lower part of the formula:
$\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 * \sum_{i=1}^{n}(y_i - \bar{y})^2}$
standardizes the covariance to both variances

```
cov(muensingen$FL, muensingen$Length) /
  sqrt(var(muensingen$FL) * var(muensingen$Length))
```

```
## [1] 0.9124495
```

covariance (cov) / square root (sqrt) of Variance Footlength * variance Length

or simpler:

```
cor(muensingen$FL, muensingen$Length)
```

```
## [1] 0.9124495
```

Correlation: coefficient of determination [1]

Specifies how much of the variation of the dependent variable is explained by the variation of the independent variable.

Example: to what percentage is the foot length explained by the fibula length?

Determination coefficient r^2= r ^2 ;-)

Our example: r = 0.9124495, r^2 = 0.832564

83.2564018% of the variation in foot length is explained by the length of the fibula!

Attention: "explained" does not necessarily mean causal connection!

Correlation test

It correlates, but does it also correlate significantly?

Test against a normally distributed error distribution with Pearson's correlation coefficient (the "normal" correlation coefficient)

The variables should be distributed normally (check with ks.test or shapiro.test)

```
shapiro.test(muensingen$FL)
```

```
##
##     Shapiro-Wilk normality test
##
## data:  muensingen$FL
## W = 0.63595, p-value = 2.37e-05
```

```
shapiro.test(muensingen$Length)
```

```
##
##     Shapiro-Wilk normality test
##
## data:  muensingen$Length
## W = 0.80529, p-value = 0.0024
```

```
# OK, in our example it is not the case.
# If it would be, we could do:

cor.test(muensingen$FL,muensingen$Length)
```

```
##
##     Pearson's product-moment correlation
##
## data:  muensingen$FL and muensingen$Length
## t = 8.6363, df = 15, p-value = 3.314e-07
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.7691077 0.9683922
## sample estimates:
##       cor
## 0.9124495
```

Correlation: least-squares method exercise

Correlation between number of millstones and number of cereal grains (Shennan example)

The number of cereal grains and millstones is given in different Neolithic settlements. Indicate how strongly the variables correlate with each other, how much of the variation of the millstones is explained by the cereal grains and whether the correlation is significant!

File: cereal_processing.csv

Correlation of ordinally scaled variables

If, as is often the case, we have no measurement data, or they are not normal distributed.

Measures for the correlation of ordinally scaled data (rank correlation):

Kendall's τ (tau)

Spearmans ρ (rho)

Example according to Shennan: Size of settlement and quality of soil

	poor	medium	good	Sum
small	15	7	2	24
medium	6	11	4	21
big	7	7	8	22
Sum	28	25	14	67

Kendall's τ (tau) [1]

Calculation over the ranks

Prerequisites: Two at least ordinally scaled variables of a random sample

Idea: With a perfect correlation, all large settlements are located on the good soils, all medium ones on the medium and all small ones on the bad.

The calculation is based on possible pairings of values whose ranks are compared to each other.

If both x and y values are smaller for a pairing than is the case at the comparison pair, the result is a concurrent pairing (with both have the same ranking).

If the x value is greater for a pairing, but the y value is smaller, then it's a discordant pair.

Kendall's τ (tau) [2]

Concurrent ranks

	poor	medium	good	Sum
small	15 (a)	7 (d)	2 (g)	24
medium	6 (b)	11 (e)	4 (h)	21
big	7 (c)	7 (f)	8 (i)	22
Sum	28	25	14	67

All settlements in cell (a) can be combined with all settlements in e,f,h,i so that both soil quality and settlement size are greater in a than in e,f,h,i.

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

Kendall's τ (tau) [3]

Concurrent ranks

	poor	medium	good	Sum
small	15 (a)	7 (d)	2 (g)	24
medium	6 (b)	11 (e)	4 (h)	21
big	7 (c)	7 (f)	8 (i)	22
Sum	28	25	14	67

All settlements in cell b can be combined with all settlements in f,i, so that both soil quality and settlement size in a are greater than in f,i.

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

Kendall's τ (tau) [4]

Concurrent ranks

	poor	medium	good	Sum
small	15 (a)	7 (d)	2 (g)	24
medium	6 (b)	11 (e)	4 (h)	21
big	7 (c)	7 (f)	8 (i)	22
Sum	28	25	14	67

All settlements in cell d can be combined with all settlements in h,i, so that both soil quality and settlement size in a are greater than in h,i.

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

Kendall's τ (tau) [5]

Concurrent ranks

	poor	medium	good	Sum
small	15 (a)	7 (d)	2 (g)	24
medium	6 (b)	11 (e)	4 (h)	21
big	7 (c)	7 (f)	8 (i)	22
Sum	28	25	14	67

All settlements in cell e can be combined with all settlements in i, so that both soil quality and settlement size in a are greater than in i.

pairings: e*i= 11*8=88

Kendall's τ (tau) [6]

Concurrent ranks

	poor	medium	good	Sum
small	15 (a)	7 (d)	2 (g)	24
medium	6 (b)	11 (e)	4 (h)	21
big	7 (c)	7 (f)	8 (i)	22
Sum	28	25	14	67

The number of pairings with concurrent ranks is therefore the sum of the individual possible pairings.

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

Kendall's τ (tau) [7]

Discordant Ranks

	poor	medium	good	Sum
small	15 (a)	7 (d)	2 (g)	24
medium	6 (b)	11 (e)	4 (h)	21
big	7 (c)	7 (f)	8 (i)	22
Sum	28	25	14	67

All settlements in cell g can be combined with all settlements in b,c,e,f, so that soil quality is worse, but settlement size is larger than in b,c,e,f.

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

Kendall's τ (tau) [7]

Discordant Ranks

	poor	medium	good	Sum
small	15 (a)	7 (d)	2 (g)	24
medium	6 (b)	11 (e)	4 (h)	21
big	7 (c)	7 (f)	8 (i)	22
Sum	28	25	14	67

All settlements in cell h can be combined with all settlements in c,f, so that soil quality is worse, but settlement size larger than in c,f.

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

Kendall's τ (tau) [8]

Discordant Ranks

	poor	medium	good	Sum
small	15 (a)	7 (d)	2 (g)	24
medium	6 (b)	11 (e)	4 (h)	21
big	7 (c)	7 (f)	8 (i)	22
Sum	28	25	14	67

All settlements in cell d can be combined with all settlements in b,c, so that soil quality is worse, but settlement size larger than in b,c.

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

Kendall's τ (tau) [9]

Discordant Ranks

	poor	medium	good	Sum
small	15 (a)	7 (d)	2 (g)	24
medium	6 (b)	11 (e)	4 (h)	21
big	7 (c)	7 (f)	8 (i)	22
Sum	28	25	14	67

All settlements in cell e can be combined with all settlements in c, so that soil quality is poorer, but settlement size is larger than in c.

pairings: e*c=11*7=77

Kendall's τ (tau) [10]

Discordant Ranks

	poor	medium	good	Sum
small	15	7	2	24
medium	6	11	4	21
big	7	7	8	22
Sum	28	25	14	67

The number of pairings with discordant ranks is therefore the sum of the individual possible pairings.

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

Kendall's τ (tau) [11]

Calculating τ:

$$\tau_c = \frac{C-D}{\frac{1}{2}*n^2*\frac{m-1}{m}} \; with \; m = min(n_{row}, n_{col})$$

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

$$\tau_c = \frac{712-286}{\frac{1}{2}*67^2*\frac{3-1}{3}}$$

$$\tau_c = \frac{426}{1496.3}$$

$$\tau_c = 0.285$$

in R:

```
soil <- read.csv2("soilsites.csv", row.names = 1)
cor.test(soil$size, soil$soil_quality, method = "kendall")
```

```
##
##     Kendall's rank correlation tau
##
## data:  soil$size and soil$soil_quality
## z = 2.6372, p-value = 0.008359
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##       tau
## 0.2902363
```

Attention: there is no calculation for Kendall's tau c, only for Kendall's tau b in R. Therefore, the data must be raw, not a contingency table.

To consider:

Correlation is not automatically a causal relationship!

Example: The well-known rattling stork example

The decrease of storks correlates with the decrease of births in Switzerland... causal connection?

Often it is hidden complex third variables that influence two correlating variables, e.g. the changes in modern society, which influence both the decline of storks and births.

More funny correlations at http://www.tylervigen.com/spurious-correlations.

https://commons.wikimedia.org/wiki/File:Storch_bringt_Baby.JPG