Two or more variables (one group)

Correlation and regression
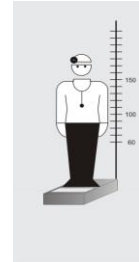
The relationship between two variables.

Method for estimating the relationship between variables.

Correlation of two variables

Example:
Is there any relationship between the height and weigth?
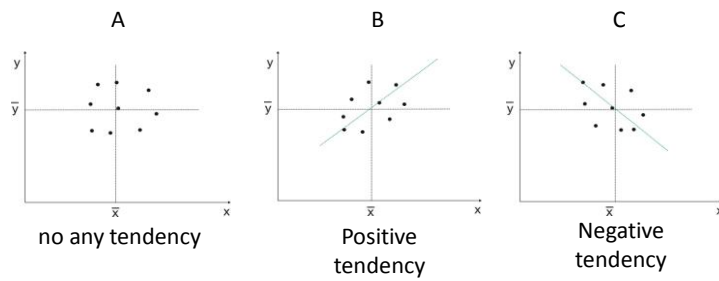
Experiment:

Data pairs:

No.	Height (cm)	Weight (kg)
1	150	61
2	170	70
3	166	75
4	174	70
5	180	72
6	155	50
7	172	65
8	161	59
9	177	81

Graphic representation

E.g.: height is the **x** and weigth is the **y**.

Possible situations:

A

no any tendency

B

Positive tendency
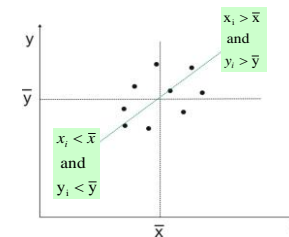
C

Negative tendency

Covariance

$$Q_{xy} = \sum_i \left[(x_i - \bar{x}) \cdot (y_i - \bar{y}) \right]$$

$$\text{cov}(x, y) = \frac{Q_{xy}}{n-1}$$
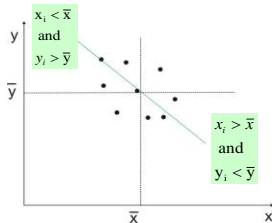
(n: no. of elements)

Positive tendency:

$x_i > \bar{x}$ and $y_i > \bar{y}$

$x_i < \bar{x}$ and $y_i < \bar{y}$

Frequently:

if $x_i < \bar{x}$ then $y_i < \bar{y}$
or $x_i > \bar{x}$ then $y_i > \bar{y}$

Consequence: $Q_{xy} > 0$.

Slide 1
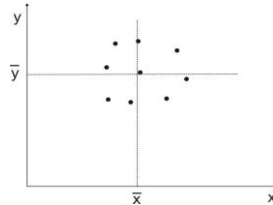
Negative tendency:

$x_i < \overline{x}$ and $y_i > \overline{y}$

$x_i > \overline{x}$ and $y_i < \overline{y}$

Frequently:

if $x_i < \overline{x}$ then $y_i > \overline{y}$

or $x_i > \overline{x}$ then $y_i < \overline{y}$

Consequence: $Q_{xy} < 0$.

No tendency:

The y values are independent from the x-values!

Consequence: $Q_{xy} = 0$.
(if n=∞)

Pearson's correlation coefficient

$$r = \frac{\text{cov}(x, y)}{s_x \cdot s_y} = \frac{Q_{xy}}{\sqrt{Q_x \cdot Q_y}}$$

Possible range for r:

$$-1 \leq r \leq 1$$

Covariance divided by the squareroot of the product of the standard deviation of the two variables (= standardized covariance).
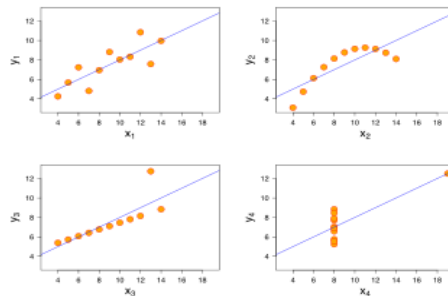The measure of the relationship.

***In the population*:**

r = 0 no correlation,

$r \neq 0$ correlation (strength is proportional to the actual value of r.)

Dependency on distribution

r-value are equal to each other in these figures!

Pearson's r value is sensitive for the skewness of the distribution and for the extreme values. Characterizes well the relationship in the case of normal distribution.

http://en.wikipedia.org/wiki/Anscombe%27s_quartet

Coefficient of determination

r^2

The coefficient of the determination tells us how strong is the relationship. Expresses how much percent of the variability of the y values may be accounted by the variability of the independent variable or variables.

Correlation *t*-test

Calculated **r** is the estimation of the **r** in the population. This fluctuates around the theoretical value.
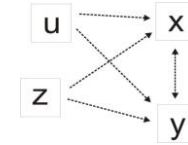(e.g. r_{calc} = 0.1 ?)

H$_0$: r = 0! $\longrightarrow$ $t = r\sqrt{\dfrac{n-2}{1-r^2}}$ $\longrightarrow$ d.f.: n - 2

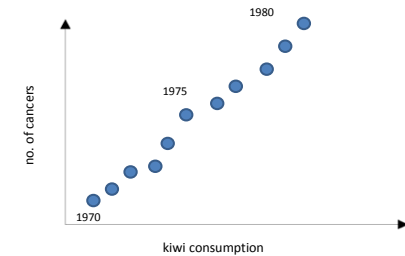**_Decision_**: based on t-value. Look previous cases!

Condition: at least one of the variable has normal distribution.

Interpretation of the correlation

Not necessary being direct causality. (But may be!)
In the background there may be quantities, effects that influence both measured variables.

Example:
There is positive correlation,
but we cann't suppose causality.

1980
1975
1970
no. of cancers
kiwi consumption

Non-normal distribution or ordinal data

Example: blood pressure measurements.
Relationship between the two methods.

The distribution is skewed, non-normal.
Pearson's r is false.

Spearman's rank-correlation

Example: diastolic pressure.
(only a few cases!)

$$r_s = \frac{\sum\left(R_x - \bar{R}_x\right)\left(R_y - \bar{R}_y\right)}{\sqrt{\sum\left(R_x - \bar{R}_x\right)^2 \cdot \sum\left(R_y - \bar{R}_y\right)^2}}$$

case	Cuff	rank	finger	rank
1	80	4.5	58	1
2	65	2	79	5
3	70	3	66	2
4	80	4.5	93	6
5	60	1	75	4
6	82	6	71	3
...				

R_x: rank of the x variable
R_y: rank of the y variable.

The test and the decision are the same as in the case of Pearson's r.

Linear regression

If the variables have normal distribution, the relationship is linear, and we can describe with straightline.

y=ax+b

(x_i, y_i)

h

$\bar{y}$

$\bar{x}$

The regression model:
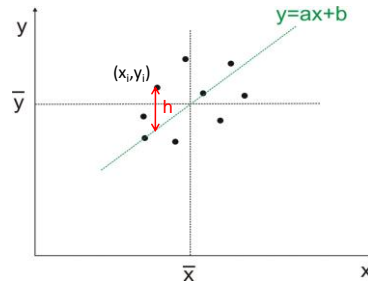(to predict the y values)

$$y_i = ax_i + b + h_i$$

In regression analysis :

$$y_i = b_0 + bx_i + h_i$$

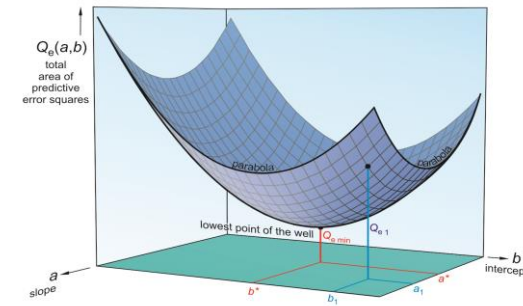y: dependent variable
x: independent (explanatory) variable
h_i: error term = $y_i - (ax_i + b)$.
    (the difference between the actual value and the predicted value.)

Least-squares method

$$Q_e = \sum_i h_i^2 = \sum_i (y_i - (ax_i + b))^2$$

The x_i and y_i are measured values. Unknown values are the a and b!

$Q_e(a,b)$
total area of predictive error squares

parabola

lowest point of the well $Q_{e\,min}$ $Q_{e\,1}$

a slope

b intercept

b^* b_1 a_1 a^*

Which is the best straightline?

Q_e has minimum! →

$$a^* = \frac{Q_{xy}}{Q_{xx}} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$b^* = \bar{y} - a^* \bar{x}$$

Prediction of insulin sensitivity by BMI.

r^2: coefficient of determination. How much part of the variance of y is explained by x.

indep.	regr. coeff.	st. error	t	p	decision
BMI	-0.077	0.018	-4.25	0.0011	significant
r^2	0.6				

Multilinear regression

Two or more independent variables.

The regression model:
(to predict the y values)
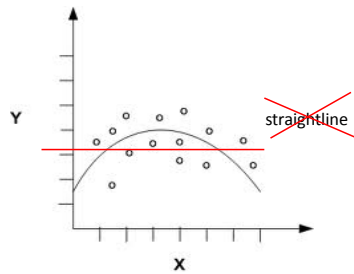
$$y_i = b_0 + \sum_j b_j x_{j,i} + h_i$$

Predict the insulin sensitivity (y)!
Conclusion: only the BMI influences the sensitivity. About the 64% of the variation of the sensitivity explained by the BMI.

indep.	regr. coeff.	st. error	t	p	decision
Age	-0.0045	0.0041	-1.09	0.3	not sign.
BMI	-0.068	0.02	-.344	0.0055	significant
r^2	0.639				

Non-linear regression

On the base of the model (e.g.):

Polinomial: $\quad y = b_0 + \sum_i b_i x^i + h$

Exponential: $\quad y = hab^x$

Power: $\quad y = hax^b$

Y

straightline

X