# Two or more variables (one group)

Correlation and regression

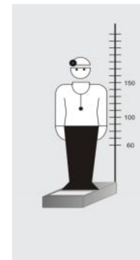
The relationship between two variables.

Method to estimate the relationship between variables.

# Correlation of two variables

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

**Experiment:**




Data pairs:

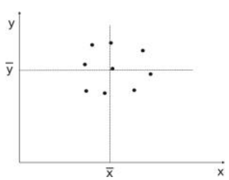
| No. | Height (cm) | Weight (kg) |
|---|---|---|
| 1 | 150 | 61 |
| 2 | 170 | 70 |
| 3 | 166 | 75 |
| 4 | 174 | 70 |
| 5 | 180 | 72 |
| 6 | 155 | 50 |
| 7 | 172 | 65 |
| 8 | 161 | 59 |
| 9 | 177 | 81 |

# Graphic representation

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

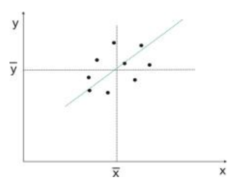
Possible situations:

A



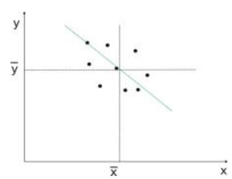

no any tendency

B




Positive tendency

C




Negative tendency

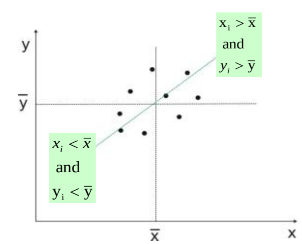
# Covariance

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

$$\mathrm{cov}(x, y) = \frac{Q_{xy}}{n-1}$$

(n: no. of elements)

Positive tendency:

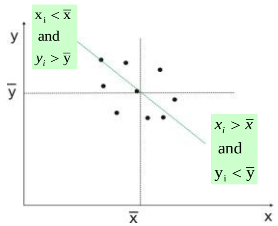



**Frequently:**

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

**<u>Consequence</u>**: $Q_{xy} > 0$.

Negative tendency:



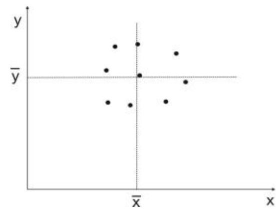

Frequently:

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

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

**<u>Consequence</u>**: $Q_{xy} < 0$.

No tendency:




The y values are independent from the x-values!

**<u>Consequence</u>**: $Q_{xy} = 0$. (if n=∞)

# Pearson's correlation coefficient

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

Possible range for r:

$$-1 \le r \le 1$$

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

***<u>In the population</u>***:

*r* = 0 no correlation,

*r* ≠ 0 correlation (strength is proportional to the actual value of *r*.)

# 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!** ⟶ $t = r\sqrt{\dfrac{n-2}{1-r^2}}$ ⟶ d.f.: n - 2

***<u>Decision</u>***: based on t-value. Look previous cases!

***<u>Condition</u>***: at least one of the variable has normal distribution.

## 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!)

| 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 |
| … | | | | |

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

$R_x$: 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

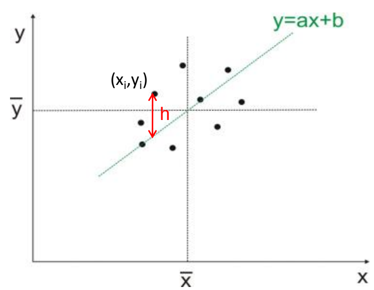
If the variables have normal distribution, the relationship is linear, and we can describe with straightline.

The regression model:
(to predict the y values)

$$y_i = ax_i + b + h_i$$

In regression analysis :

$$y_i = b_0 + bx_i + h_i$$

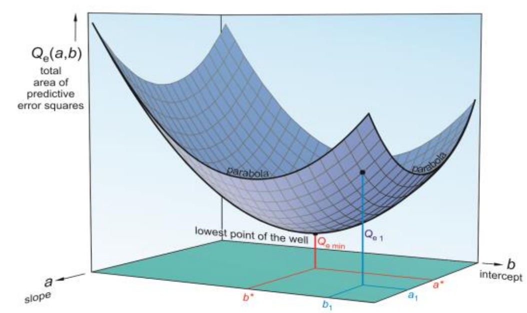



**y**: dependent variable
**x**: independent (explanatory) variable
**$h_i$**: error term = $y_i - (ax_i+b)$.
(the difference between the actual value and the predicted value.)

## 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***!



## Which is the best straightline?

$Q_e$ has minimum! → $a^* = \frac{Q_{xy}}{Q_{xx}} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$ $\quad b^* = \bar{y} - a^* \bar{x}$

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 | | | | |

## Chi-square test
## Analyzing frequency data

Example: headache

pill

Effective:
no headache

Not effective:
remaining headache

## Experiment

1. group: patients taking the ***medicine***
   - headache (*a*)
   - no headache (*b*)

2. group: patients taking the ***placebo***
   - headache (c)
   - no headache (d)

(a,b,c,d are frequency data)

## Contingency table

| | headache | no headache | Total |
|---|---|---|---|
| 1. group | a | b | a+b |
| 2. group | c | d | c+d |
| total | a+c | b+d | n |

So-called 2 x 2 table.

# Nullhypothesis

If the medicine is similar to the placebo, we expect:

$$\frac{a}{b} = \frac{c}{d} \longrightarrow a \times d = b \times c$$

***Nullhypothesis***: the medicine is similar to the placebo.

**Chi-Square test for independence.**

# Independent case

Remember: P(AB) = P(A) x P(B) if A and B are independent from each other.
(P(AB), P(A) and P(B) are estimated by relative frequencies.)

$$\frac{a}{n} \approx \frac{a+b}{n} \times \frac{a+c}{n}$$

a/n ~ P(AB) – (no effect in 1, group)
(a+b)/n ~ P(A) – (belongs to the 1. group)
(a+c)/n ~ P(B) - (no effect)

**O**bserved proportion: a/n

**E**xpected proportion: $\frac{a+b}{n} \times \frac{a+c}{n}$

transformation

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

# $\chi^2$-distribution

Shortcut formula for 2 x2 tables:

$$\chi^2 = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

***Nullhypothesis***: $\chi^2$-value is equal to 0,
the difference is due to the sampling error.

***$\chi^2$-distribution*** describes the random deviations of the $\chi^2$-value.

# Decision

Same, then in the case of *t*-distribution. We use $\chi^2$-distribution.

Expected value is 0, if the null hypothesis is true.

$p \leq \alpha$ - reject the null hypothesis else accept.

***degree of freedom***: in this special case = 1.

In general:

d.f.=(r-1)(c-1), where r – no. of rows
c – no. of columns

## Small expected frequencies

May not be used if:
1. An expected frequency is 2 or less.
2. More than 20% of the expected frequencies are less than 5.

***Fisher's exact test*** may be used.
Calculates the exact probability for the given table.

$$P = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!}$$

Remember!
n! = multiplying the integers from 1 to n.

Decision is based on the P.

## The Chi-Square test Goodness-of-Fit test

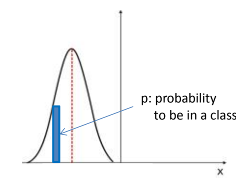
Example: testing normality of the larger diameter of the frog red blood cells.

**O**bserved frequencies:

size (μm)

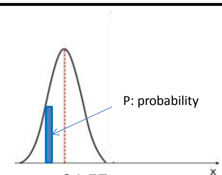
| 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | n |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 4 | 10 | 9 | 20 | 26 | 27 | 37 | 42 | 48 | 53 | 45 | 39 | 35 | 17 | 18 | 10 | 7 | 5 | 450 |

Nullhypothesis ($H_0$):
Data has normal distribution. Calculate the average and the sd from the sample!
Calculate expected frequency from the normal distribution!
in a class = np (see figure)




## Chi-Square test

avg = 24.57 μm;
sd = 3.62 μm




$n_i$ ~ 450 x P

μ = 24.57 μm

**E**xpected frequencies:

| 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2.8 | 5.1 | 8.9 | 14.2 | 21 | 29 | 37 | 44 | 48 | 49 | 46.4 | 41 | 33 | 25 | 18 | 11 | 6.9 | 7.2 |

**O**bserved frequencies:

| 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | n |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 4 | 10 | 9 | 20 | 26 | 27 | 37 | 42 | 48 | 53 | 45 | 39 | 35 | 17 | 18 | 10 | 7 | 5 | 450 |

Degree of freedom = m – b – 1
m: no. of classes. (in example = 18)
b: no. of parameters (in example = 2)

Calculation:
p = 0.96

We accept the nullhypothesis.

## Chi-Square test Test for homogenity

**Example**: The frequency of wearing glasses is the same in the groups of girls and boys or not?

$H_0$: There is no difference.
(independent!)

P(**W**ith) ~ 76/200; P(**B**oys) ~ 97/200
Independent case:
P(W and B) = P(W) x P(B)
expected freq. ~ n x (PW and B)
= 200 * 76/200 * 97/200 ~ 36.9

**O**bserved frequencies

| | with | without | |
|---|---|---|---|
| boys | 48 | 49 | 97 |
| girls | 28 | 75 | 103 |
| | 76 | 124 | 200 |

Expected frequencies

| | with | without | |
|---|---|---|---|
| boys | 36.9 | 60.1 | 97 |
| girls | 39.1 | 63.9 | 103 |
| | 76 | 124 | 200 |

# Calculation

$$\chi^2=\frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}=\frac{200\cdot(48\cdot 75-49\cdot 28)^2}{76\cdot 124\cdot 97\cdot 103}$$

$\chi^2 \approx 10.5$ d.f. = 1 $p \approx 0.001$

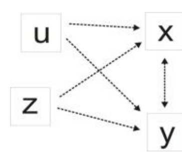
Decision:
We reject the nullhypothesis. There is significant difference between boys and girls.

# Conditions for tests

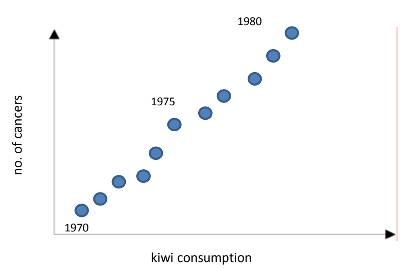
| test | condition |
|---|---|
| One-sample t-test | One group, one variable, normal distribution |
| Two-sample t-test | Two independent groups, one variable, normal distribution, the standard deviations may be the same in the groups |
| ANOVA | 3 or more independent groups, one variable, normal distribution |
| Sign test | One group, numerical or ordinal quantity |
| Wilcoxon's signed rank-test | One group, numerical or ordinal quantity |
| Mann-Whitney U-test | Two independent groups, numerical or ordinal quantity |
| Kruskall-Wallis test | 3 or more groups, numerical quantity |
| Pearson's correlation test | One group, two variables, normal distribution |
| Spearman's correlation test | One group, two variables, numerical or ordinal quantity |
| Chi-Square test (independency) | Two or more groups, frequency data |
| Chi-Square test (homogenity) | Two or more groups, frequency data |
| Chi-Square test (fit) | One group, known distribution, frequency data |

# Interpretation of the correlation

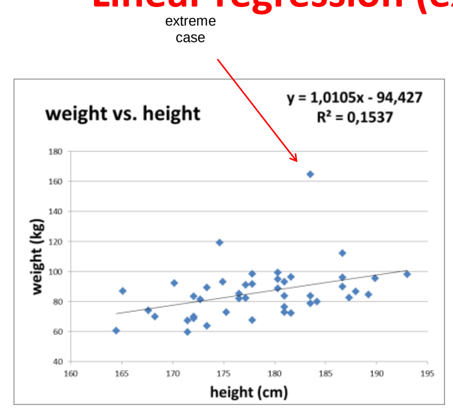
Not necessary being direct causality. (But may be!)
In the background there may be quantities, effects that influence both measured variables.




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



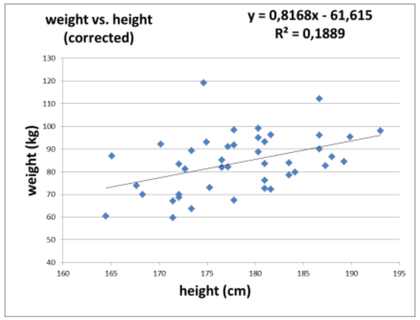

# Linear regression (example 1)




correlation
t-test

| | |
|---|---|
| n | 44 |
| r | 0,3920 |
| t | 2,761 |
| p | 0,85% |

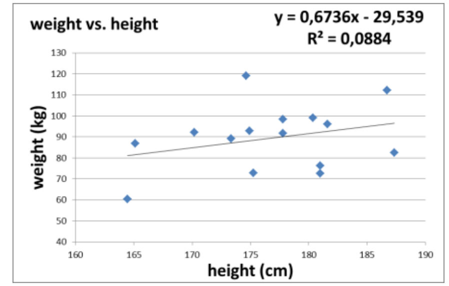
# Linear regression (without extreme case)




correlation
t-test

| | |
|---|---|
| n | 43 |
| r | 0,4347 |
| t | 3,090 |
| p | 0,36% |

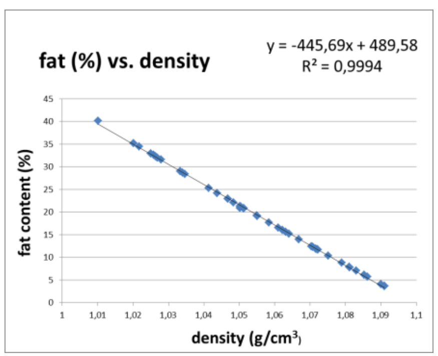
# Linear regression (effect of n)




correlation
t-test

| | |
|---|---|
| n | 15 |
| r | 0,2973 |
| t | 1,123 |
| p | 28,20% |

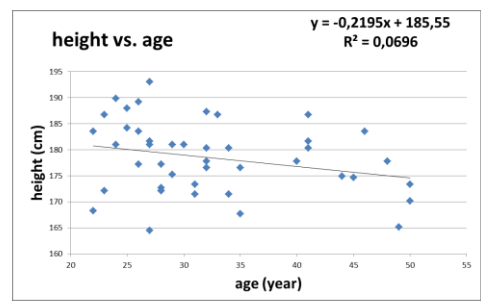
# Linear regression (example 2)




correlation
t-test

| | |
|---|---|
| n | 44 |
| r | -0,9997 |
| t | -271,242 |
| p | 9,33E-70 |

# Linear regression (example 3)




correlation
t-test

| | |
|---|---|
| n | 44 |
| r | -0,2638 |
| t | -1,772 |
| p | 8,36% |

# Multilinear regression

Two or more independent variables.

The regression model:
(to predict the y values)

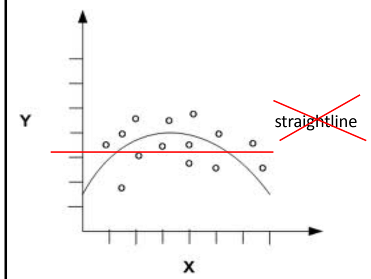
$$y_i = b_0 + \sum_j b_j x_{j,i} + h_i$$

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

Exponential: $y = hab^x$

Power: $y = hax^b$