How dependent is one measure of the other, and how to model this dependence?

Correlation and its coefficients (Kendall's tau, Spearman's rho, Pearson's r). Linear regression.

Katalin Mády

Pázmány Summer Course in Linguistics for Linguistics Students, Piliscsaba

27 August, 2013

Relevance of sample size for hypothesis testing

Relevance of n:
The critical values for the confidence interval $p = 0.95$ are calculated as
lower value: $-1.96 * se + \bar{x}$

upper value: $-1.96 * se + \bar{x}$

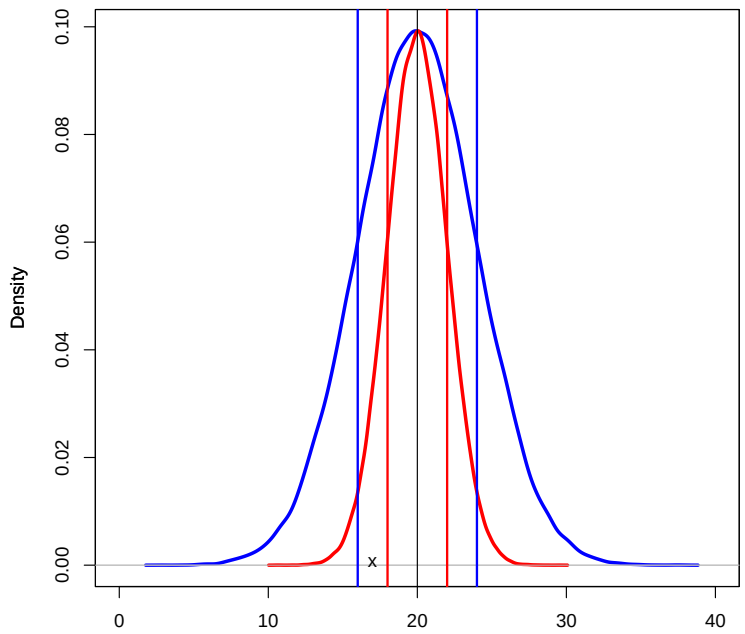
The standard error is calculated with the formula

$$se = \frac{s}{\sqrt{n}}$$

If n is small, the standard error will be larger, since it has $\sqrt{n}$ in the denominator $\Rightarrow$ a larger standard error leads to a larger critical value within which H_0 cannot be rejected.

Example

Two populations with $\mu = 20$. $se = 1$ for population 1, $se = 2$ for population 2 $\rightarrow$
lower critical value for population 1:20-1*1.96 = 18.04, for population 2: 20-2*1.96 = 16.08.
Does a sample with $\bar{x} = 17$ belong to this population?



Population 1: $17 < 18.04 \Rightarrow H_0$ can be rejected.
Population 2: $17 > 16.08 \Rightarrow H_0$ cannot be rejected.

Relationship between variables

The relationship between two dependent variables or between one dependent and one independent variable where any variable is at least ordinal can be expressed by correlation coefficients.
Possible correlation types:

- While a is increasing, b is also increasing.
- While a is increasing, b is decreasing.
- There is no relationship between the two variables.

Possible interpretations:

1. The variable a depends on the variable b. Typical example: years of education and yearly income.
2. Both variables depend on a third, unknown or uninvestigated variable. In the above example, family background might be such a factor.

Strength of correlation

Range between -1 and $+1$.

- If there is a strong positive relationship between two variables, the correlation coefficient is close to $+1$.
- If there is a strong negative relationship between two variables, the correlation coefficient is close to -1.
- If there is no relationship, the correlation coefficient is close to 0.

Correlation coefficients: Pearson's r for metrical data, Spearman's ρ and Kendall's τ for ordinal data.

Covariance

If two measures depend on each other, then the deviation of x from the mean should result in a deviation of y from the mean in a positive or negative direction.

Calculation of covariance:

1. The deviation of each data point from the mean is calculated.
2. The deviations for corresponding data points are multiplied $\rightarrow$ all distances will be positive $\sim$ SD.
3. All distances are summed.

$$\sigma_x, y = \frac{1}{n-1} \sum_{k=1}^{n} (x_k - \vec{x}) * (y_k - \vec{y})$$

The lower limit of covariance is 0. There is no upper limit, i.e. it is infinite.

Standardisation of covariance

Problem with covariance: the value depends on sample size and standard deviation.

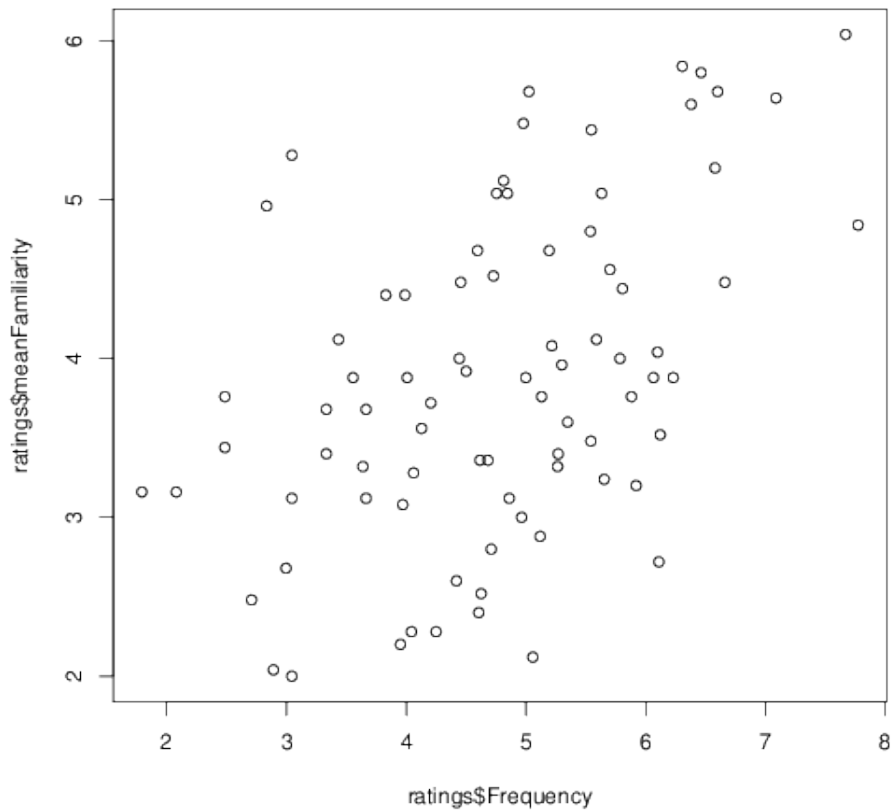
The origin of coordinate system is the mean. Covariance is then divided by the SD of the two variables:
Pearson's r

$$r = \frac{\sum_{k=1}^{n} (x_k - \vec{x}) * (y_k - \vec{y})}{\sum_{k=1}^{n} (x_k - \vec{x})^2 * \sum_{k=1}^{n} (y_k - \vec{y})^2}.$$

Pearson's r is the correlation coefficient for normally distributed metrical data.

Example

English native speakers were asked how familiar they were with 81 animals and plants. They gave scores on a 7 steps scale. Are the names of plants and animals with which they are more familiar also more frequent?



Example

Frequency of words is a parametric measure since it is a count. The scale of familiarity scores are in theory ordinal, but such scales are accepted as metrical measures if the distance between each two neighbouring score is equal.

The image shows that frequency and familiarity are correlated to some extent, and that the correlation is positive.

Pearson's $r = 0.48$

This is a strong positive correlation.

Strength of correlation

Interpretation of r

+.70 or higher: very strong positive relationship
+.40 to +.: strong positive relationship
+.30 to +.39: moderate positive relationship
+.20 to +.29: weak positive relationship
+.01 to +.19: no or negligible relationship
-.01 to -.19: no or negligible relationship
-.20 to -.29: weak negative relationship
-.30 to -.39: moderate negative relationship
-.40 to -.69: strong negative relationship
-.70 or higher: very strong negative relationship

Spearman's *Rho* (ρ)

If one or both variables are ordinal, Pearson's r is not applicable. Instead, Spearman's ρ can be used.

Procedure: values in both variables are sorted in ascending order, the deviations (d) are squared and summed.

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2-1)}$$

Example

IQ and number of hours per week people spend with watching TV.

IQ	h	rank x	rank y	d	d^2
86	1	1	1	0	0
97	20	2	6	-4	16
99	28	3	8	-5	25
100	27	4	7	-3	9
101	50	5	10	-5	25
103	29	6	9	-3	9
106	7	7	3	4	16
110	17	8	5	3	9
112	6	9	2	7	49
113	12	10	4	6	36

ρ=-29/165=-0.17575757

Kendall's τ

Another correlation coefficient for ordinal data.

- Advantage: it is more reliable for samples with small numbers and many ties than ρ.
- Disadvantage: its square cannot be interpreted as the coefficient of determination (see below).

Procedure: goodness of rank of elements divided by the number of possible pairs.
Proversion (P): number of elements of y that fit into expected rank.
Inversion (I): deviation from the expected rank.

Two sets: $a = [1, 2, 3, 4], b = [2, 1, 3, 4]$
number of proversions and inversions: $P_{b1} = 2, I_{b1} = 1, P_{b2} = 2, P_{b3} = 1$

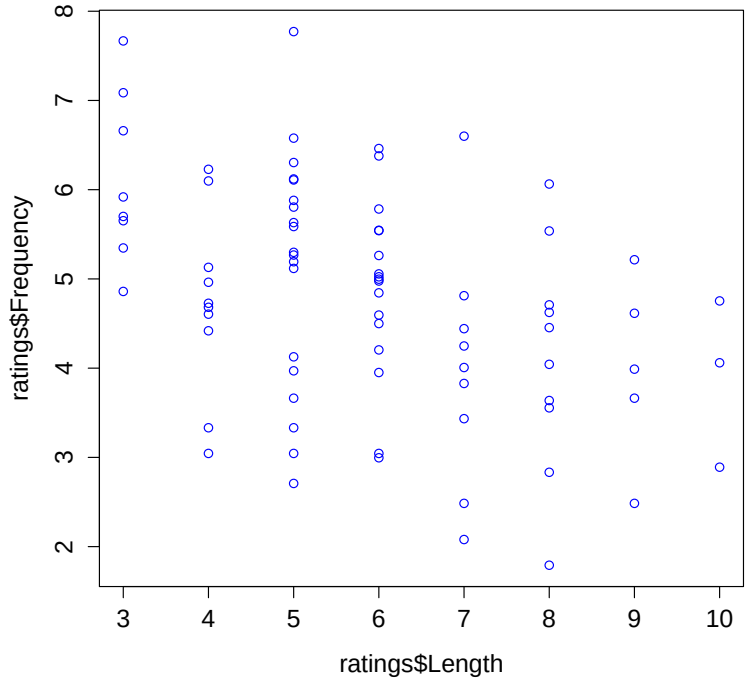
$$\tau = \frac{P - I}{\frac{n(n-1)}{2}} = \frac{5-1}{6} = 0.6666666667$$

Regression analysis

- Looking for the function between one or more continuous explanatory data or independent variables and a dependent variable.
- Only applicable to metrical scales.
- There are univariate and multivariate regressions.
- The regression function can be linear (of 1st order) or of higher order (a parabola or a polynom).

Linear regression

Correlation between length of animal and plant names and their frequency.



Correlation coefficients: $r = -0.43$, $\rho = -0.43$, $\tau = -0.31$.

Regression line

How do we find the line that expresses the correlation best?

Regression line

How do we find the line that expresses the correlation best?

Calculation of correlation coefficients: sum of products of distances from mean,
that is

$$\sigma_x, y = \frac{1}{n-1} \sum_{k=1}^{n} (x_k - \bar{x}) * (y_k - \bar{y})$$

Regression line

How do we find the line that expresses the correlation best?

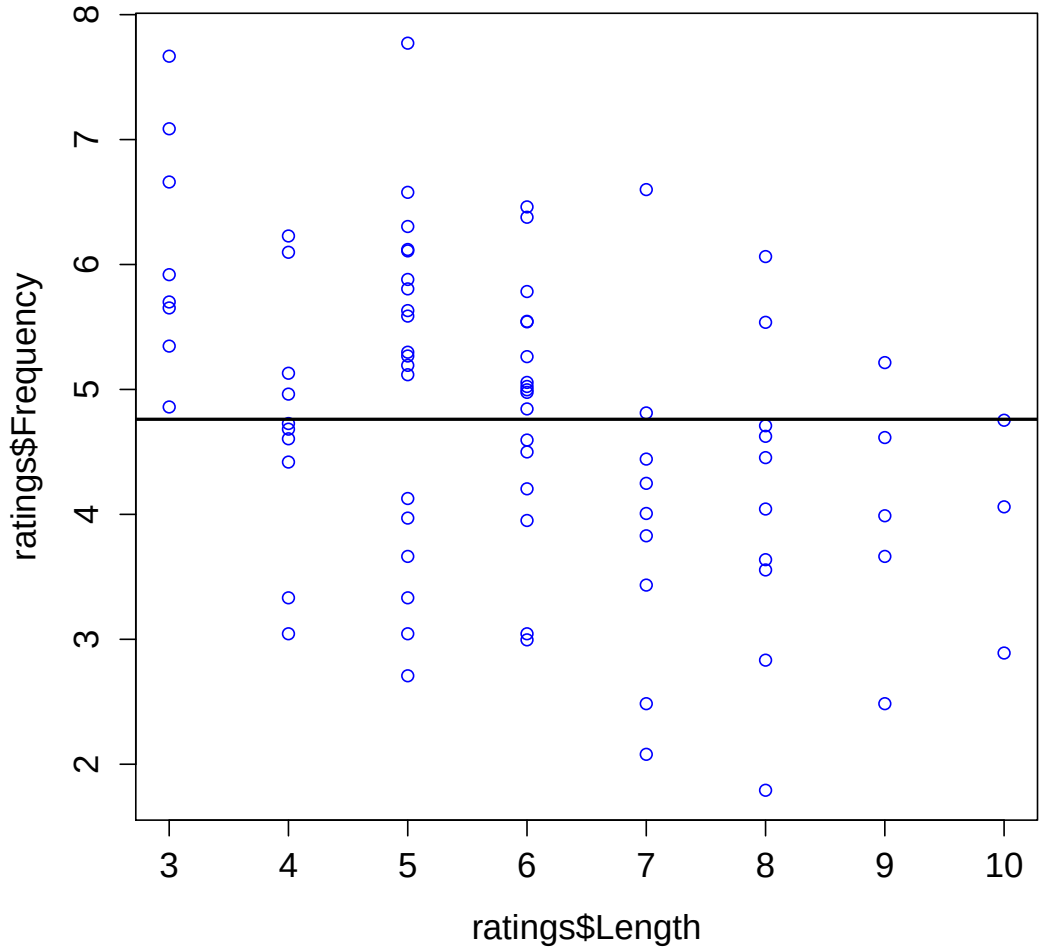
Calculation of correlation coefficients: sum of products of distances from mean,
that is

$$\sigma_x, y = \frac{1}{n-1} \sum_{k=1}^{n} (x_k - \bar{x}) * (y_k - \bar{y})$$

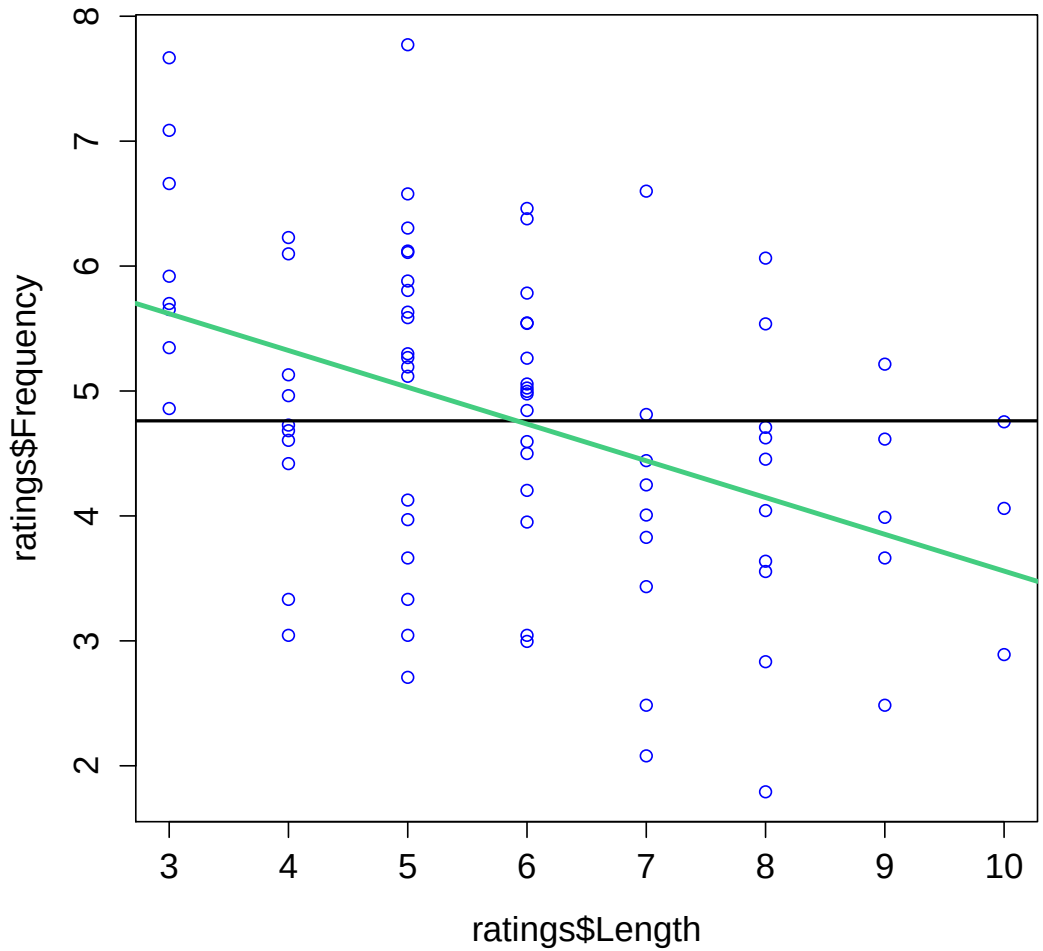
Difference from correlation tests: we are not interested in the relationship of variables, but in the effect that the independent variable x has on the dependent variable y. This should be expressed by a function.

Ordinary least squares

The residual of each y value, i.e. deviation from the mean, is calculated. Then their squares are summed.



We are looking for the line from wich the squared vertical distance of y values (residual) is smallest.



Function of line

$$y = a + bx$$

Function of line

$$y = a + bx$$

Regression coefficients:

a: intercept of line on y-axis.
b: steepness of line.

Ordinary Least Square (OLS):

$$OLS = \sum_{k=1}^{n} (y_i - (a + bx_i))^2 = min$$

Coefficient of determination

Regression line is the best approximation to describe the function.

Coefficient of determination

Regression line is the best approximation to describe the function.

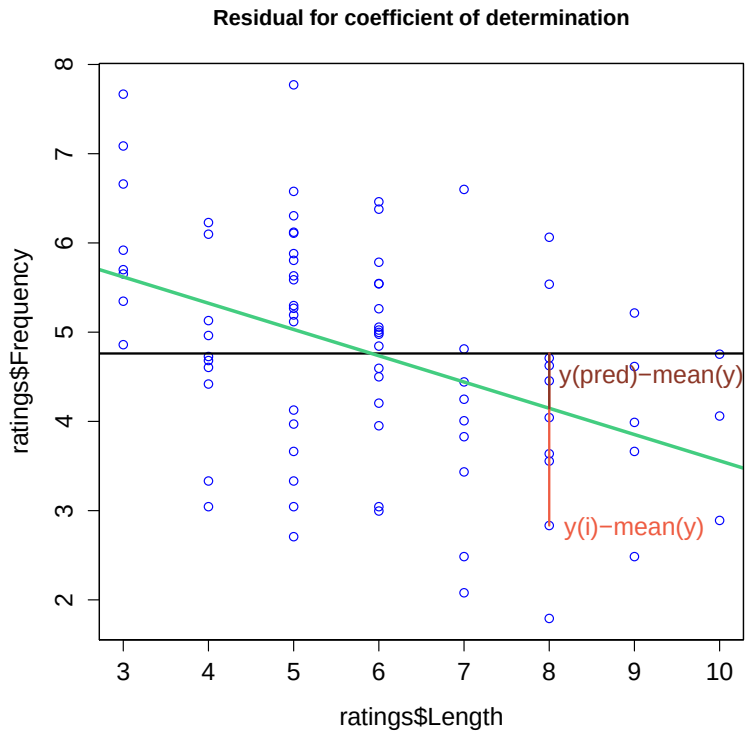
How can we calculate how good the approximation is (goodness of fit)?

Coefficient of determination

Regression line is the best approximation to describe the function.

How can we calculate how good the approximation is (goodness of fit)?

By the sum of squares of residuals, but hier not based on the deviation from the mean, but on the distance from the mean plus the regression line.



Coefficient of determination

Calculation:

$$d = R^2 = \frac{SS_R}{SS_e/(n-2)}$$

where SS = sum of squares, R: regression, e: error, and $(n-2)$ is necessary to calculate the standard deviation.

If data are normally distributed, then the value is identical with the square of the correlation coefficient r^2.

Interpretation: what proportion of the variation of y can be explained by x.

Relevance: the amount of variation in the y that can be explained by x is smaller than the correlation coefficient (since it is its square root).

In the above example: $r = 0.43$, $r^2 = 0.18$. Thus even if data are strongly correlated, only 18% of the variation in frequency can be explained based on word length.