DESCRIPTIVE STATISTICS

Dr Alina Gleska

Institute of Mathematics, PUT

April 20, 2018

The unit of the population - the pair (X, Y). The correlation does not depend on the order of variables but this order (and the distinguishing which variable is independent and which is dependent) is very important in the regression analysis.

At first we have to decide if there is any bond between those variables - we do not make any correlation analysis without checking it.

Aims of the correlation analysis:

- checking if there is any dependence between variables,
- what is the shape of this relationship (linear, nonlinear),
- what is its strength,
- what is its direction.

Measures of correlations depend on the shape of the relation so at first we assess the shape, and then the power and the direction of this relation.

We distinguish two types of dependence:

- functional - for each value of X we have only one value of Y;
- statistical (stochastic) - for each value of X we have many values of Y.

Graphical presentation of the data:

- correlation series – bivariate series
- the scatterplot – a plot of the points (x_i, y_i) in the plane
- the correlation table – the table for grouped data

Graphical presentation of the data:

- correlation series – bivariate series
- the scatterplot – a plot of the points (x_i, y_i) in the plane
- the correlation table – the table for grouped data

Graphical presentation of the data:

- correlation series – bivariate series
- the scatterplot – a plot of the points (x_i, y_i) in the plane
- the correlation table – the table for grouped data

A key feature in a scatterplot is the association, or trend between X and Y.

- the strength of relationship – we estimate it basing on the dispersion of variables; higher dispersion means lower strength of relationship;
- the direction of correlation – when higher values of X tend to be paired with higher values of Y, these two values have a positive association; when higher values of X tend to be paired with lower values of Y, these two values have a negative association. If higher X values are paired with low or with high Y values equally often, there is no association.
- the shape of relationship – a mathematical function between variables; a linear or a nonlinear one.

A key feature in a scatterplot is the association, or trend between X and Y.

- the strength of relationship – we estimate it basing on the dispersion of variables; higher dispersion means lower strength of relationship;
- the direction of correlation – when higher values of X tend to be paired with higher values of Y, these two values have a positive association; when higher values of X tend to be paired with lower values of Y, these two values have a negative association. If higher X values are paired with low or with high Y values equally often, there is no association.
- the shape of relationship – a mathematical function between variables; a linear or a nonlinear one.

A key feature in a scatterplot is the association, or trend between X and Y.

- the strength of relationship – we estimate it basing on the dispersion of variables; higher dispersion means lower strength of relationship;
- the direction of correlation – when higher values of X tend to be paired with higher values of Y, these two values have a positive association; when higher values of X tend to be paired with lower values of Y, these two values have a negative association. If higher X values are paired with low or with high Y values equally often, there is no association.
- the shape of relationship – a mathematical function between variables; a linear or a nonlinear one.

A key feature in a scatterplot is the association, or trend between X and Y.

- the strength of relationship – we estimate it basing on the dispersion of variables; higher dispersion means lower strength of relationship;
- the direction of correlation – when higher values of X tend to be paired with higher values of Y, these two values have a positive association; when higher values of X tend to be paired with lower values of Y, these two values have a negative association. If higher X values are paired with low or with high Y values equally often, there is no association.
- the shape of relationship – a mathematical function between variables; a linear or a nonlinear one.

Correlation table

For discrete or qualitative variables:

Variable X	Variable Y				Total
	y_1	y_2	$\dots$	y_ℓ	
x_1	n_{11}	n_{12}	$\dots$	$n_{1\ell}$	$n_{1\bullet}$
x_2	n_{21}	n_{22}	$\dots$	$n_{2\ell}$	$n_{2\bullet}$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
x_k	n_{k1}	n_{k2}	$\dots$	$n_{k\ell}$	$n_{k\bullet}$
Total	$n_{\bullet 1}$	$n_{\bullet 2}$	$\dots$	$n_{\bullet\ell}$	n

Correlation table

For continuous variables:

Variable X	Variable Y				Total
	$[y_0, y_1)$	$[y_1, y_2)$	$\dots$	$[y_{\ell-1}, y_\ell]$	
$[x_0, x_1)$	n_{11}	n_{12}	$\dots$	$n_{1\ell}$	$n_{1\bullet}$
$[x_1, x_2)$	n_{21}	n_{22}	$\dots$	$n_{2\ell}$	$n_{2\bullet}$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
$[x_{k-1}, x_k]$	n_{k1}	n_{k2}	$\dots$	$n_{k\ell}$	$n_{k\bullet}$
Total	$n_{\bullet 1}$	$n_{\bullet 2}$	$\dots$	$n_{\bullet \ell}$	n

Marginal distribution of a variable X

Marginal distribution is a distribution of one variable alone.

Variable X	Variable Y				frequency
	$[y_0, y_1)$	$[y_1, y_2)$	$\dots$	$[y_{\ell-1}, y_\ell]$	
$[x_0, x_1)$	n_{11}	n_{12}	$\dots$	$n_{1\ell}$	$n_{1\bullet}$
$[x_1, x_2)$	n_{21}	n_{22}	$\dots$	$n_{2\ell}$	$n_{2\bullet}$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
$[x_{k-1}, x_k]$	n_{k1}	n_{k2}	$\dots$	$n_{k\ell}$	$n_{k\bullet}$
Total	$n_{\bullet 1}$	$n_{\bullet 2}$	$\dots$	$n_{\bullet \ell}$	n

Marginal distribution of a variable Y

Variable X	Variable Y				Total
	$[y_0, y_1)$	$[y_1, y_2)$	$\dots$	$[y_{\ell-1}, y_\ell)$	
$[x_0, x_1)$	n_{11}	n_{12}	$\dots$	$n_{1\ell}$	$n_{1\bullet}$
$[x_1, x_2)$	n_{21}	n_{22}	$\dots$	$n_{2\ell}$	$n_{2\bullet}$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
$[x_{k-1}, x_k]$	n_{k1}	n_{k2}	$\dots$	$n_{k\ell}$	$n_{k\bullet}$
frequency	$n_{\bullet 1}$	$n_{\bullet 2}$	$\dots$	$n_{\bullet\ell}$	n

Conditional distribution

Conditional distribution is a distribution of one variable under the condition that the second one has got a concrete value.
The distribution of X under the condition $y \in [y_1, y_2)$

Variable X	Variable Y				Total
	$[y_0, y_1)$	$[y_1, y_2)$	$\dots$	$[y_{\ell-1}, y_\ell]$	
$[x_0, x_1)$	n_{11}	n_{12}	$\dots$	$n_{1\ell}$	$n_{1\bullet}$
$[x_1, x_2)$	n_{21}	n_{22}	$\dots$	$n_{2\ell}$	$n_{2\bullet}$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
$[x_{k-1}, x_k)$	n_{k1}	n_{k2}	$\dots$	$n_{k\ell}$	$n_{k\bullet}$
Total	$n_{\bullet 1}$	$n_{\bullet 2}$	$\dots$	$n_{\bullet \ell}$	n

Conditional distribution

The distribution of Y under the condition $x \in [x_1, x_2)$

Variable X	Variable Y				Total
	$[y_0, y_1)$	$[y_1, y_2)$	$\dots$	$[y_{\ell-1}, y_\ell]$	
$[x_0, x_1)$	n_{11}	n_{12}	$\dots$	$n_{1\ell}$	$n_{1\bullet}$
$[x_1, x_2)$	n_{21}	n_{22}	$\dots$	$n_{2\ell}$	$n_{2\bullet}$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
$[x_{k-1}, x_k]$	n_{k1}	n_{k2}	$\dots$	$n_{k\ell}$	$n_{k\bullet}$
Total	$n_{\bullet 1}$	$n_{\bullet 2}$	$\dots$	$n_{\bullet\ell}$	n

Pearson's correlation coefficient

If the relationship between variables is linear we use Pearson's correlation coefficient. Its formula:

$$r = \frac{cov(X, Y)}{S_X \cdot S_Y},$$

where $cov(X, Y)$ is the covariance of the variables X and Y, and S_X - the standard deviation of the variable X, S_Y - the standard deviation of the variable Y.

Pearson's correlation coefficient

Properties:

- the coefficient r is a symmetric measure, e.g. its value does not depend whether we check the dependence the variable X on the variable Y or the opposite case,
- $r \in [-1, 1]$,
- the sign of r, which is coincide with the sign of the covariance, informs us about the direction of the correlation:
  - (i) $r > 0$ – positive correlation (an increase of the value of one variable implies an increase of the mean value of the second variable),
  - (ii) $r < 0$ – negative correlation (an increase of the value of one variable implies a decrease of the mean value of the second variable),
- the strength of the correlation:

Pearson's correlation coefficient

- the strength of the correlation:
  - (i) $|r| \in [0, 0.2)$ – a very weak linear relationship,
  - (ii) $|r| \in [0.2, 0.4)$ – a weak linear relationship,
  - (iii) $|r| \in [0.4, 0.6)$ – a moderate linear relationship,
  - (iv) $|r| \in [0.6, 0.8)$ – a strong linear relationship,
  - (v) $|r| \in [0.8, 1)$ – a very strong linear relationship,
  - (vi) $|r| = 1$ – a functional relationship (the points lie on the straight line).

Pearson's correlation coefficient

The formula for the Pearson's correlation coefficient for the individual bivariate data:

$$cov(X, Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{xy}$$

Remark: the covariance has got units (for example: years $\cdot$ cm), but they are not interpreted. Its sign is very important because tells us about the direction of the correlation.

Pearson's correlation coefficient

The standard deviations S_X and S_Y we calculate using the formulas:

$$S_X = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}x_i^2 - \overline{x}^2}$$

and

$$S_Y = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}y_i^2 - \overline{y}^2}.$$

Pearson's correlation coefficient

The formula for the Pearson's correlation coefficient for the grouped bivariate data:

$$cov(X,Y) = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{k}(x_i^0 - \overline{x})(y_j^0 - \overline{y})n_{ij} = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{k}x_i^0 y_j^0 n_{ij} - \overline{xy},$$

where x_i^0 – the center of the $i-th$ interval of the variable X (in the discrete case we write x_i instead of x_i^0 where x_i is the $i-th$ category of the variable X), similarly for y_j^0, and n_{ij} is the frequency from the correlation table lying in the $i-th$ row and $j-th$ column.

Pearson's correlation coefficient

The standard deviations S_X and S_Y we calculate using the formulas:

$$S_X = \sqrt{\frac{1}{n}\sum_{i=1}^{r}(x_i^0 - \overline{x})^2 n_{i\bullet}} = \sqrt{\frac{1}{n}\sum_{i=1}^{r}(x_i^0)^2 n_{i\bullet} - \overline{x}^2},$$

where $n_{i\bullet}$ – the sum of frequencies in rows (the marginal frequency in the distribution of the variable X)
and

$$S_Y = \sqrt{\frac{1}{n}\sum_{j=1}^{k}(y_j^0 - \overline{y})^2 n_{\bullet j}} = \sqrt{\frac{1}{n}\sum_{j=1}^{k}(y_j^0)^2 n_{\bullet j} - \overline{y}^2},$$

where $n_{\bullet j}$ – the sum of frequencies in columns (the marginal frequency in the distribution of the variable Y).

Spearman's rank correlation coefficient

Spearman's rank correlation coefficient is used for estimating the direction and the strength of the relationships in two situations:

- when variables are non-measureable but we can order them,
- when variables are measureable and the number of their categories is finite.

Spearman's rank correlation coefficient

At first we have to order values of observations and give them the numbers from 1 to n. This process is called the ranking of the variables.
Remark 1: The ranking must be the same for both variables (in the ascending order of the values or in the descending order of the values of both variables).
Remark 2: Identical values are usually each assigned fractional ranks equal to the average of their positions in the ascending order of the values, which is equivalent to averaging over all possible permutations.

Spearman's rank correlation coefficient

Variable X	Variable Y	Rank X	Rank Y
x_1	y_1	1	1
x_2	y_2	2	3
x_3	y_3	3,5	4
x_4	y_4	3,5	2
x_5	y_5	5	5
$\vdots$	$\vdots$	$\vdots$	$\vdots$

Spearman's rank correlation coefficient

Formula:

$$r_s = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} d_i^2,$$

where n is the total number of the observation of one variable, d_i – the difference between the two ranks of each observation.

Spearman's rank correlation coefficient

PROPERTIES:

- Spearman's rank correlation coefficient r_s is a symmetric measure,
- Spearman's rank correlation coefficient is robust (while Pearson's correlation coefficient is not robust and depends on outliers),
- $r_s \in [-1, 1]$,

Spearman's rank correlation coefficient

- the sign tells us about the direction of the relationship:
  - (i) $r_s > 0$ – positive correlation (correlation between two variables will be high when observations have a similar rank),
  - (ii) $r_s = 1$ – when observations have an identical rank,
  - (iii) $r_s = 0$ – no correlation,
  - (iv) $r_s < 0$ – negative correlation (correlation between two variables will be low when observations have a dissimilar rank),
  - (v) $r_s = -1$ – when observations have a fully oposed rank,

Spearman's rank correlation coefficient

- the strength of the correlation:
  - (i) $|r_s| \in [0, 0.2)$ – a very weak relationship,
  - (ii) $|r_s| \in [0.2, 0.4)$ – a weak relationship,
  - (iii) $|r_s| \in [0.4, 0.6)$ – a moderate relationship,
  - (iv) $|r_s| \in [0.6, 0.8)$ – a strong relationship,
  - (v) $|r_s| \in [0.8, 1)$ – a very strong relationship.

EXAMPLE.

Correlation analysis - qualitative variables

If we are not able to give ranks variables we have to use so called contingency coefficients. They consist of:

- φ – Yule's coefficient,
- T – Czuprow's coefficient,
- V – Cramer's coefficient,
- P – Pearson's coefficient.

Correlation analysis - qualitative variables

All contingency coefficients are based on χ^2 statistics:

$$\varphi = \sqrt{\frac{\chi^2}{n}}, \quad T = \sqrt{\frac{\chi^2}{n\sqrt{(r-1)(k-1)}}},$$

$$V = \sqrt{\frac{\chi^2}{n \cdot \min(r-1, k-1)}}, \quad P = \sqrt{\frac{\chi^2}{\chi^2 + n}}.$$

Correlation analysis - qualitative variables

Properties of φ, T, V, P coefficients:

- they are symmetric,
- they are non-negative,
- they say nothing about the direction of the correlation,
- $\varphi, T, V, P \in [0, 1]$,
- they tell us about the strength of the correlation:
  - (i) $0 - 0.2$ – a very weak relationship,
  - (ii) $0.2 - 0.4$ – a weak relationship,
  - (iii) $0.4 - 0.6$ – a moderate relationship,
  - (iv) $0.6 - 0.8$ – a strong relationship,
  - (v) $0.8 - 1$ – a very strong relationship.

χ^2 statistics

χ^2 statistics we calculate as follows:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{k} \frac{(n_{ij} - n_{ij}^t)^2}{n_{ij}^t},$$

where n_{ij} – the empirical frequencies, n_{ij}^t – the theoretical frequencies.

χ^2 statistics

The theoretical frequencies we calculate as follows:

$$n_{ij}^t = \frac{n_{i\bullet} \cdot n_{\bullet j}}{n},$$

where $n_{i\bullet}$ and $n_{\bullet j}$ are marginal frequencies, and n – the total number of observations.
EXAMPLE