MS-C1620 Statistical inference

6 Correlation and dependence

Jukka Kohonen

Department of Mathematics and Systems Analysis
School of Science
Aalto University

Academic year 2020–2021
Period III–IV

Contents

Independence and dependence

- Two random variables/experiments are independent if the result of one does not in any way help us predict the result of the other.
- More formally, the random variables x and y are independent if for all (suitable) sets A, B we have,

$$\mathbb{P}(x \in A \mid y \in B) = \mathbb{P}(x \in A).$$

- If the above does not hold, the random variables x, y are called dependent.
- Saying that two random variables are dependent does not, however, give any indication on the type of dependence or how dependent they are.

Dependence in statistics

In statistics, the dependence of random variables is usually of major interest.

- The dependence between unemployment rate and (growth of) GDP in Finland, election promises, etc.
- The dependence between alcohol consumption and alcohol price, income level, availability of alcohol, warning labels, etc.
- The dependence between incidence of lung cancer and smoking (duration, amount of cigarettes) etc.

Linear dependence

- The simplest form of dependence is linear dependence.
- If the random variables x and y satisfy,

$$y = ax + b,$$

  for some constants $a, b \in \mathbb{R}, a \neq 0$, then the variable y is a linear transformation of the variable x and the random variables x and y are said to be (completely) linearly dependent.
- Linear dependence between two variables can be measured, for example, using the (Pearson) correlation coefficient.

Pearson correlation coefficient

- Let $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ be i.i.d. observations of a bivariate random variable (x, y).
- Then the sample covariance,

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}),$$

  estimates the population covariance $E[(x - E[x])(y - E[y])] = \sigma_{xy}$,
- and the sample Pearson correlation coefficient,

$$\hat{\rho} = \hat{\rho}(x, y) = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}},$$

  estimates the Pearson correlation coefficient

$$\rho = \rho(x, y) = \frac{\sigma_{xy}}{\sigma_x \sigma_y}.$$

Interpreting the Pearson correlation

- Pearson correlation coefficient numerically measures the linear dependence between two random variables. The coefficient is always in the interval $[-1, 1]$ and attains the values ± 1 if and only if $y = ax + b$ for some $a, b \in \mathbb{R}, a \neq 0$.

- If the variables x and y are independent, then the Pearson correlation coefficient $\rho(x, y) = 0$.
- However, the contrary does not hold. That is, $\rho(x, y) = 0$ does not imply the independence of x and y (take, for example, x standard normal and $y = x^2$).
- *Thus there are also other forms of dependence than linear dependence.*

Example 1

Examples of data exhibiting linear dependence of various degrees:

Example 2

Examples of data exhibiting complete linear dependency (correlation coefficients equal to ± 1) (in the middle one $\rho(x, y)$ cannot be computed due to division by zero):

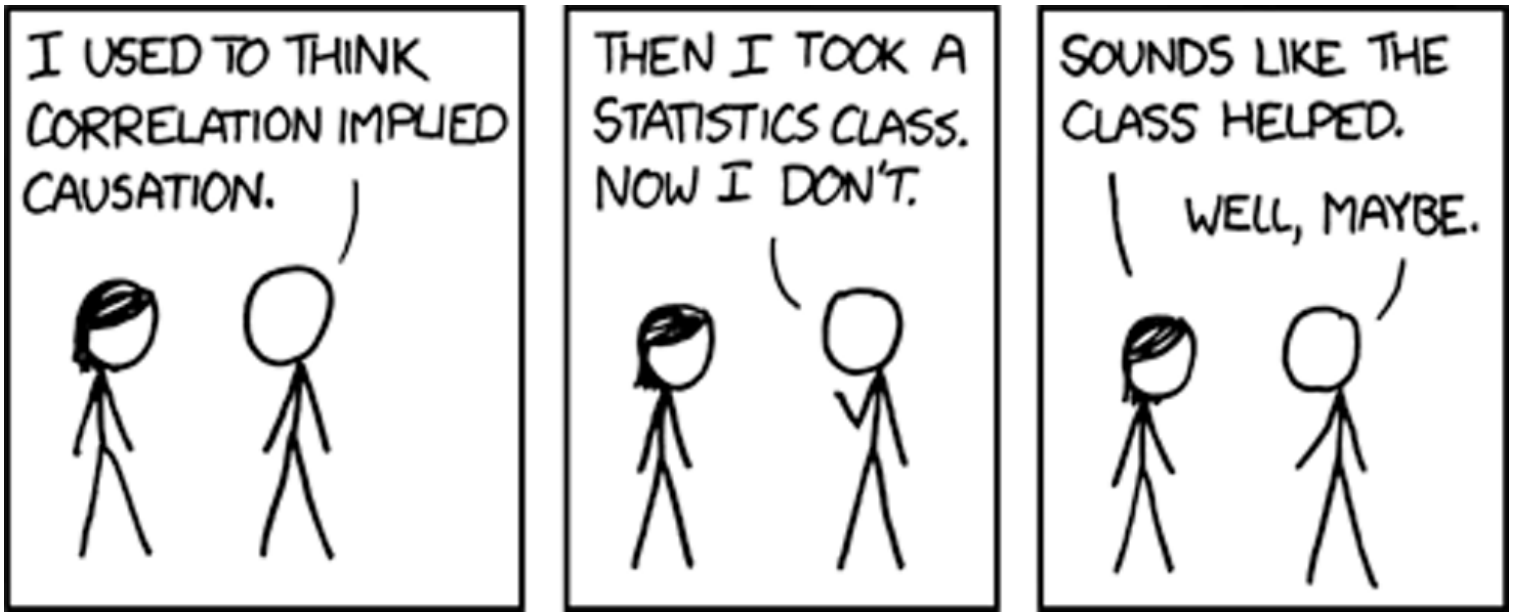
Example 3

Examples of weird shaped data all having approximately zero linear dependency (but still having clear dependency of some other form):

Check also http://guessthecorrelation.com/ and https://www.autodeskresearch.com/publications/samestats

Correlation vs. causation

Note that showing that two things are dependent reveals nothing about their causal relation (which one caused the other).



Check out: http://www.tylervigen.com/spurious-correlations.

Three different warnings!

1. Sample correlation only *estimates* the "true" correlation. So even if true distribution has *zero* correlation, a small sample usually has nonzero. $\longrightarrow$ Try confidence intervals or tests.

   E.g. I just rolled two dice, 30 times each. The correlation coefficient was -0.12.

2. Correlation could be small or zero, but still there could be *nonlinear* dependence. $\longrightarrow$ See end of this lecture about monotonic dependence.

3. Correlation between X and Y could be big, (so big that it is not just random noise in sample), but it does not mean that X *causes* Y.

   It could be the other way round. Or it could be a third variable that causes both to be big at the same time. Finding causality is very important, but requires stronger tools than we have on this course. At least a high correlation can be taken as a *hint* of possible causality.

Bivariate normal distribution

The bivariate normal distribution is an extension of the normal distribution into two dimensions.

Let (x, y) have the bivariate normal distribution. Then its marginal distributions are normal distributions with the expected values $\mu_x, \mu_y \in \mathbb{R}$ and the variances $\sigma_x^2, \sigma_y^2 > 0$.

In addition to these, the bivariate normal distribution has the parameter $\rho \in [-1, 1]$, the Pearson correlation between the marginals.

The probability density function of a bivariate normal distribution is,

$$f(x, y) = \frac{1}{2\pi\sqrt{1-\rho^2}\sigma_x\sigma_y} \cdot$$

$$\exp\Big(-\frac{1}{2(1-\rho^2)}\Big(\frac{(x-\mu_x)^2}{\sigma_x^2} - 2\rho\frac{(x-\mu_x)}{\sigma_x}\frac{(y-\mu_y)}{\sigma_y} + \frac{(y-\mu_y)^2}{\sigma_y^2}\Big)\Big).$$

Parametric confidence interval for Pearson correlation

Let $(x_1, y_1), (x_2, y_2) \ldots, (x_n, y_n)$ be an iid. sample from the bivariate normal distribution and denote,

$$\tanh(t) = \frac{e^{2x} - 1}{e^{2x} + 1}, \quad \operatorname{arctanh}(t) = \frac{1}{2} \log\left(\frac{1+t}{1-t}\right)$$

Then $\operatorname{arctanh}(\hat{\rho})$ is (for large n) approximately normally distributed and this can be used to derive an approximate $100(1-\alpha)$ level confidence interval for ρ:

$$\left[\tanh\left(\operatorname{arctanh}(\hat{\rho}) - z_{\alpha/2}\frac{1}{\sqrt{n-3}}\right), \tanh\left(\operatorname{arctanh}(\hat{\rho}) + z_{\alpha/2}\frac{1}{\sqrt{n-3}}\right)\right],$$

where $z_{\alpha/2}$ is the $(1-\alpha/2)$-quantile of the standard normal distribution.

Note that the above confidence interval makes the assumption of bivariate normal distribution.

Note also that the confidence intervals for the Pearson correlation provided by statistical software are almost always based on normality assumption.

Non-parametric confidence interval for Pearson correlation

If the iid. bivariate sample $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ does not come from the bivariate normal distribution, a non-parametric alternative to the previous parametric confidence interval is given by the bootstrap.

Non-parametric confidence interval for Pearson correlation

1. Pick new sample of n pairs from the observed pairs $(x_1, y_1), (x_2, y_2) \ldots, (x_n, y_n)$ with replacement, such that new pairs are selected one-by-one and the selected pair is each time "returned back" to the original sample (the same pair can be selected multiple times).
2. Estimate the Pearson correlation coefficient for the new sample formed in the previous step.
3. Repeat the previous steps B times.
4. After the replications, order the B estimates from the smallest to the largest.
5. A $100(1-\alpha)\%$ confidence interval is now obtained by choosing the $\lfloor B \times (\alpha/2) \rfloor$ ordered estimate as the lower endpoint and the $\lfloor B \times (1-\alpha/2) \rfloor$ ordered estimate as the upper endpoint.

One-sample test for Pearson correlation

The one-sample test for the Pearson correlation coefficient compares the Pearson correlation to a given constant under the assumption of bivariate normality.

One-sample test for Pearson correlation, assumptions

Let $(x_1, y_1), (x_2, y_2) \ldots, (x_n, y_n)$ be an iid. random sample from a bivariate normal distribution.

One-sample test for Pearson correlation, hypotheses

$$H_0 : \rho = \rho_0 \qquad H_1 : \rho \neq \rho_0.$$

One-sample test for Pearson correlation

One-sample test for Pearson correlation, test statistic

- The test statistic,

$$z = \frac{\operatorname{arctanh}(\hat{\rho}) - \operatorname{arctanh}(\rho_0)}{\sqrt{\frac{1}{n-3}}},$$

  follows under H_0 (for large n) approximately the standard normal distribution.
- The expected value of z under H_0 is 0 and if the value of the test statistic has **large absolute value**, evidence against the null hypothesis H_0 is found.

Two-sample test for Pearson correlation

The two-sample test for Pearson correlation compares the Pearson correlation coefficients of two independent samples under the assumption of bivariate normality.

Two-sample test for Pearson correlation, assumptions

Let $(x_1, y_1), (x_2, y_2) \ldots, (x_n, y_n)$ be an iid. random sample from a bivariate normal distribution with the Pearson correlation ρ_1 and let $(z_1, w_1), (z_2, w_2) \ldots, (z_m, w_m)$ be an iid. random sample from a bivariate normal distribution with the Pearson correlation ρ_2. Furthermore, let the two samples be independent.

Two-sample test for Pearson correlation, hypotheses

$$H_0 : \rho_1 = \rho_2 \qquad H_1 : \rho_1 \neq \rho_2.$$

Two-sample test for Pearson correlation

Two-sample test for Pearson correlation, test statistic

- The test statistic,

$$z = \frac{\operatorname{arctanh}(\hat{\rho}_1) - \operatorname{arctanh}(\hat{\rho}_2)}{\sqrt{\frac{1}{n-3} + \frac{1}{m-3}}},$$

  follows under H_0 (for large n and m) approximately the standard normal distribution.
- The expected value of z under H_0 is 0 and if the value of the test statistic has **large absolute value**, evidence against the null hypothesis H_0 is found.

Parametric significance test for Pearson correlation

Often it is of interest to assess whether the Pearson correlation differs statistically significantly from zero (no correlation). Under the assumption of bivariate normality, an approximate test for this can be carried out using the one-sample test for Pearson correlation.

However, an exact test can also be performed.

Parametric significance test for Pearson correlation, assumptions

Let $(x_1, y_1), (x_2, y_2) \ldots, (x_n, y_n)$, be an i.i.d. random sample from a bivariate normal distribution.

Parametric significance test for Pearson correlation, hypotheses

$$H_0 : \rho = 0 \qquad H_1 : \rho \neq 0.$$

Parametric significance test for Pearson correlation

Parametric significance test for Pearson correlation, test statistic

- The test statistic,

$$t = \sqrt{n-2} \cdot \frac{\hat{\rho}}{\sqrt{1-\hat{\rho}^2}},$$

  follows under H_0 the t_{n-2}-distribution.
- The expected value of z under H_0 is 0 and if the value of the test statistic has **large absolute value**, evidence against the null hypothesis H_0 is found.

Non-parametric significance test for Pearson correlation

A non-parametric alternative to the parametric significance test is given by a permutation test.

Permutation test for Pearson correlation, assumptions

Let $(x_1, y_1), (x_2, y_2) \ldots, (x_n, y_n)$, be an i.i.d. random sample from a bivariate distribution.

Permutation test for Pearson correlation, hypotheses

$$H_0 : \rho = 0 \qquad H_1 : \rho \neq 0.$$

Permutation test for Pearson correlation

Let $\hat{\rho}$ be the Pearson correlation of the original sample. The probability of obtaining a value equally or more deviating than $\hat{\rho}$ under the null hypothesis can be estimated using a permutation test as follows.

1. Form n new pairs $(x_1, y_1^*), (x_2, y_2^*) \ldots, (x_n, y_n^*)$ from the original observed pairs $(x_1, y_1), (x_2, y_2) \ldots, (x_n, y_n)$, such that $y_1, y_2, \ldots, y_n$ are permuted randomly and each original y_j is used only once in the new sample.
2. Estimate Pearson correlation using the new sample $(x_1, y_1^*), (x_2, y_2^*) \ldots, (x_n, y_n^*)$.
3. Repeat the steps 1 and 2 several times.
4. Estimate the probability of obtaining a value equally or more deviating than $\hat{\rho}$ under the null hypothesis using the generated distribution of estimates. That is, calculate the percentage of the generated estimates in that have **absolute value** greater than $|\hat{\rho}|$.

Permutation test for Pearson correlation

- The permutation test is based on the idea that the permuted samples $(x_1, y_1^*), (x_2, y_2^*) \ldots, (x_n, y_n^*)$ do not exhibit correlation (as the pairs are chosen randomly) and if $\hat{\rho}$ differs a lot from the typical correlation coefficient of a permuted sample, we can conclude that the original sample exhibits significant correlation.
- More accurate estimate can be achieved using a permutation test without simulation. In an exact permutation test, all possible $n!$ sample combinations are used, and the probability of obtaining the value $\hat{\rho}$ or more extreme value under the null hypothesis is estimated exactly using all $n!$ correlation coefficients.

Contents

Monotonic dependence

- A more flexible form of dependency is given by monotonic dependence.
- If the random variables x and y satisfy

$$y = g(x),$$

  where g is a monotonic (increasing or decreasing) function, then y is a monotonic transformation of x and the random variables x and y are said to be (completely) monotonically dependent.
- The monotonic dependence between two random variables can be measured using, for example, Spearman's rank correlation coefficient.

Spearman's rank correlation coefficient

- Let $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ be i.i.d. observations of a bivariate random variable (x, y).
- Let $R(x_i)$ denote the rank of the observation x_i in the sample $x_1, x_2, \ldots, x_n$ and let $R(y_i)$ denote the rank of the observation y_i in the sample $y_1, y_2, \ldots, y_n$.
- Then Spearman's rank correlation coefficient $\rho_S(x, y)$ is the Pearson's correlation coefficient calculated from the ranks.

Spearman's rank correlation coefficient

- Spearman's rank correlation coefficient measures the monotonic dependence between two random variables. The coefficient is always in the interval $[-1, 1]$ and (in case of no repeating data values) attains the absolute value 1 if and only if $y = g(x)$ for some monotonic function g.
- If the variables x and y are independent, then the Spearman correlation $\rho_S(x, y) = 0$ (using the same counterexample as with Pearson correlation, we see that the contrary does not again hold).
- See, link 1 and link 2 for values of the Spearman correlation for some particular data sets.

Non-unique ranks

- It is possible that some of the sample points have the same rank.
- In that case, all those points are assigned to have the median of the corresponding ranks.
- For example, if two observations have the same rank, corresponding to ranks 7 and 8, then both are assigned to have rank 7.5. If three observations have the same rank, corresponding to ranks 3, 4, and 5, then each is assigned to have rank 4.

Confidence intervals for Spearman correlation

Confidence intervals for Spearman's rank correlation coefficient can be estimated using the bootstrap.

Significance tests for Spearman correlation

Significance test for Spearman correlation can be conducted non-parametrically either via a permutation test or through the following.

Significance test for Spearman correlation, assumptions

Let $(x_1, y_1), (x_2, y_2) \ldots, (x_n, y_n)$, be an i.i.d. random sample from a bivariate distribution.

Significance test for Spearman correlation, hypotheses

$$H_0 : \rho_S = 0 \qquad H_1 : \rho_S \neq 0.$$

Significance test for Spearman correlation

Significance test for Spearman correlation, test statistic

- The test statistic,

$$t = \sqrt{n-2} \cdot \frac{\hat{\rho}_S}{\sqrt{1-\hat{\rho}_S^2}},$$

  follows under H_0 (for large n) approximately the t_{n-2}-distribution.
- The expected value of z under H_0 is approximately 0 and if the value of the test statistic has **large absolute value**, evidence against the null hypothesis H_0 is found.