The use of correlation functions in thoracic surgery research

Lucile Gust, Xavier Benoit D’journo

Department of Thoracic Surgery and Diseases of the Oesophagus, North University Hospital, Aix-Marseille Université, France

Correspondence to: Xavier Benoit D’journo, MD, PhD. Service de Chirurgie thoracique, AP-HM, Hôpital Nord, Chemin des Bourrely, 13915 Marseille Cedex 20, France. Email: Xavier.D’journo@ap-hm.fr.

Submitted Dec 16, 2014. Accepted for publication Dec 30, 2014.
doi: 10.3978/j.issn.2072-1439.2015.01.54

View this article at: http://dx.doi.org/10.3978/j.issn.2072-1439.2015.01.54

Introduction

Correlation functions consist of a broad variety of statistical tests, used to describe the relationship between two, or more, sets of data. Those functions, whether they are parametric or non-parametric, are used in medical studies to characterize the strength of this relationship, and the direction of the relationship.

We will successively introduce parametric tests such as the Pearson correlation coefficient, the linear regression model, followed with non-parametric tests such as the Spearman and Kendall rank correlation tests. Finally we will say a few words of calibration, used to evaluate risk scores. The principle of each test will be described and illustrated with selected examples found in medical literature.

Pearson correlation coefficient

Principle

The Pearson product-moment correlation coefficient, more often known as the Pearson Correlation coefficient (also referred to as the Pearson's $r$), is a measure of the linear correlation between two variables (1,2).

For two variables $x$ and $y$, the Pearson correlation coefficient is defined as the covariance of $x$ and $y$ divided by the product of their standard deviations. The Pearson coefficient varies from $-1$ to $+1$, where negative values indicate that $y$ decrease with $x$ and positive values indicate that $y$ increase with $x$. The strength of association between the two variables is considered important (or very strong) if the coefficient ranges from 0.8 to 1, moderate (or strong) from 0.5 to 0.8, weak (or fair) from 0.2 to 0.5, and very weak (or poor) when less than 0.2. When the correlation coefficient is equal to 0, the two variables are independent from one another. When it is equal to 1 they are perfectly correlate in a positive way, whereas when it is equal to $-1$ the variables are perfectly correlate negatively (anti-correlation).

Mathematically, the correlation coefficient is expressed by the formula:

$$r = \frac{cov(x,y)}{\sqrt{var(x)var(y)}} = \frac{\sum (x_i - m_x)(y_i - m_y)}{\sqrt{\sum (x_i - m_x)^2 \sum (y_i - m_y)^2}}$$

Where $cov$ is the covariance, $var$ the variance, and $m$ the standard score of the variable.

For a given value of $r$ and the degrees of freedom of the variables, the correlation table described by Fisher and Yates allows estimation of the risk $\alpha$.

Selected literature example

Wu et al. studied the Haller index of their patients on pre-operative CT-scans and chest X-rays, before a Nuss procedure was undertaken (minimally invasive technique of surgical correction of pectus excavatum) (3). For 154 patients, they compared the numerical values of each pair of Haller Index on pre-operative chest X-rays ($x$) and CT-scan ($y$), and studied the correlation between the variables with the Pearson’s correlation coefficient, represented on a graph (Figure 1). With a correlation coefficient of 0.757, they concluded that the two chest imaging modalities had a good correlation for the Haller Index pre-operatively.

Another example of correlation can be found in the study published by D’journo et al. in 2009 (4). In this study, 94 patients with locally advanced esophageal cancer were included, with lymph node involvement in 32 patients. Metastasis was detected in 120 lymph nodes and extra-capsular extension in 49 lymph nodes. The authors used a Pearson correlation function in the first part of their results to summarize the relationship between the number of
nodes with extra-capsular lymph node involvement (y) and number of positive lymph node (x) (Figure 2). They showed a strong, significant, correlation: r=0.8; P=0.01.

**Limits**

There are limits to the use of the Pearson correlation test. It can only be used for two sets of numerical variables, one set at least with a constant variance. Moreover though the test evaluates the dependence between those variables it does not presume of the causal relationship involved.

**Linear regression**

**Principle**

Linear regression is an approach used to characterized the dependence relationship between two variables when one is the direct cause of the other (for example the link between the dose of a drug and the length of survival of a patient), when the modification of value of one variable induces as a direct consequence a modification of the other value, or when the main purpose of the analysis is the prediction of one variable from the other (1,2,5).

For a given n number of pair of two variables x and y, linear regression model can be described with a simple formula:

\[ y = x\beta + \epsilon \]

Where \( \beta \) is the regression coefficient and \( \epsilon \) represents the disturbance term or noise.

The regression coefficient can be estimated:

\[ \frac{\sum x_i y_i - (\sum x_i \sum y_i)/n}{\sum x_i^2 \times (\sum x_i)^2/n} \]

Several conditions must be met for an adequate use of linear regressions models. The relationship between the variables must be linear, the predictor variable (x) measurements are considered without error, both variables must have similar variance, and the errors of the response variable (y) must be uncorrelated to each other.

**Selected literature example**

A perfect example of linear regression can be found in the article of Arnaoutakis et al. (6). The authors compared the charges of hospital stay with the Society of Thoracic Surgeon (STS) risk scores during aortic valve replacement. Because the histogram of absolute charges revealed a non-normal distribution, they performed a logarithmic transformation to respect assumptions of linear regression. Furthermore the authors verified the variance of STS risk scores before creating the model. Finally these precautions allowed them to use a regression linear analysis to describe the relationship between log charges (y) and STS risk...
score \((x)\) during aortic valve replacement (Figure 3). The authors demonstrated a positive correlation between the two variables (correlation coefficient 0.06, 95\% CI: 0.05-0.07, \(P=0.01\)).

### Limits

Linear regression analysis can be used both for univariate or multivariate analysis. But the assumptions needed for its application must be respected. Furthermore the distribution of the variables must be monotonic. For example, if a variable \(x\) has a parabolic distribution, the relationship with another variable \(y\) if described with a straight line, used for linear regression, will be inadequate.

As with the Pearson correlation coefficient, linear regression analysis does not presume of the causal link between the studied variables.

### Non-parametric tests: Spearman and Kendall correlation

#### Spearman's rank correlation coefficient

**Principle**

Spearman’s correlation coefficient can be used to measure the dependent relationship between both continuous and discrete variables, when this relationship is not linear (as opposed to the classic Pearson correlation) (1,2). It can also be used when comparing one continuous variable and one categorical variable.

For two set of variables \(x\) and \(y\), each raw scores \(x_i\) and \(y_i\), are converted to ranks \(X_i\) and \(Y_i\). The Pearson correlation coefficient is then used on the ranked variables, following the formula:

\[
\rho = \frac{6\Sigma d_i^2}{n(n^2-1)}
\]

Where \(d_i = X_i - Y_i\) is the difference between ranks.

The coefficient \(\rho\) ranges from \(-1\) to \(+1\), in a similar way as the Pearson correlation coefficient. Accordingly the interpretation of the values is the same as well.

#### Kendall rank correlation coefficient

**Principle**

Kendall rank correlation coefficient, also called Kendall’s tau (\(\tau\)) coefficient, is also used to measure the nonlinear association between two variables (1,2,5). It is used for measured quantities, to evaluate between two sets of data the similarity of the orderings when ranked by each of their quantities.

It can be expressed with the formula:

\[
\tau = \frac{[(concor)-(discor)]}{[0.5n(n-1)]}
\]

Where \(concor\) is the number of concordant pairs and \(discor\) the number of discordant pairs.

The Kendall rank ranges from \(-1\) to \(+1\) and has a similar

---

**Figure 3** reproduced from the article of Arnaoutakis et al. (*J Thorac Cardiovasc Surg* 2011). Univariate linear regression analyses for aortic valve replacement with STS score as an independent variable and log charges used as dependent variables.

---
interpretation as both Pearson and Spearman’s correlation coefficients.

**Selected literature example**

Sinai CZ studied post-operative high-density lipoprotein cholesterol (HDL-C) levels in children undergoing the Fontan operation (surgical treatment of single ventricle defects) (8). The Kendall rank correlation coefficient was used to correlate The HDL-C levels (baseline and 24-h post-operative) with duration of post-operative chest tube drainage and total hospital length of stay. Though in this case no statistical significance was found.

Another study realized by Wang ZT et al. illustrates the use of the Kendall rank correlation coefficient (9). For 39 patients, the authors used the Kendall correlation to analyze the factors related to the percentage of volume of functional lung receiving ≥20 Gray (FLv20). They showed that the number of non-functional lung had a negative relation with the percentage of FLv20 decrease (r=-0.559, P<0.01), while the distance of planning target volume and non-functional lung center had a positive relation with the percentage of decrease FLv20 (r=0.768, P<0.01).

**Limits**

The Kendall rank correlation coefficient is used to test the strength of association between to ordinals variables, without presuming of their causal relationship.

**Calibration**

**Principle**

Medical studies often use multivariate logistic regression analysis, such as Cox logistic regression, to predict clinical outcomes and create Risk scores. These models are adequate for both categorical and continuous variables, and censored responses as well. Once the model is create, before it is used in a clinical setting, its fit with reality must be evaluated (10). Calibration is defined as the agreement between observed outcomes and predicted probabilities, and is evaluate with goodness of fit tests (5).

**Selected literature example**

Steyerberg et al., used a multivariate logistic regression analysis to identify criteria linked to 30-day mortality after esophagectomy (11) Based on these results the authors proposed a score to evaluate pre-operatively the surgical mortality of patients with esophageal cancer. In order to assess the performance of this score they studied the calibration of the score with the Hosmer-Lemeshow goodness of fit test (Figure 5). Moreover, in this study, the authors assessed the discrimination of the score. Discrimination refers to the ability to distinguish patients who will die from those who will survive. In this example it was quantified by the area under the receiver operating characteristic curve (AUC). The score evaluate in this study showed a low discrimination but a good calibration.

**Limits**

Calibration has a very specific use and aim. The goodness of fit test chosen depends of the type of explanatory variables
previously studied with the logistic regression analysis.

**Conclusions**

Correlation functions and associated statistical tools such as regression models are widely used in medical studies. Though they each have specific domains of applications, they share the limit of being unable to evaluate the causal relationship between the studied variables. Basic knowledge of correlation functions is necessary for the clinician, both to perform and understand medical studies.

**Acknowledgements**

*Disclosure:* The authors declare no conflict of interest.

**References**


**Cite this article as:** Gust L, D’Journo XB. The use of correlation functions in thoracic surgery research. *J Thorac Dis* 2015;7(3):E11-E15. doi: 10.3978/j.issn.2072-1439.2015.01.54