BMJ

BMJ 2012;345:e4483 doi: 10.1136/bmj.e4483 (Published 4 July 2012)

ENDGAMES

STATISTICAL QUESTION

Pearson's correlation coefficient

Philip Sedgwick *senior lecturer in medical statistics*

Centre for Medical and Healthcare Education, St George's, University of London, Tooting, London, UK

Researchers investigated the relation between the number of involuntary admissions (detentions) for mental disorders a year under the Mental Health Act 1983 and the number of NHS psychiatric beds each year in England. They used hospital episode statistics from 1996 to 2006 in a retrospective analysis. For each year they obtained the number of available NHS psychiatric beds—defined as those beds for patients with mental disorders or learning disabilities—and the number of involuntary admissions for mental disorders in NHS hospital and private facilities combined.[1]

It was reported that the number of NHS psychiatric beds fell in each successive year and that overall from 1996 to 2006 the number had decreased by 29%. A significant correlation existed between the number of psychiatric NHS beds each year and the combined number of involuntary admissions for mental disorders to NHS and private facilities under the Mental Health Act 1983 (Pearson correlation coefficient r=−0.94 (P<0.001)).

Which of the following statements, if any, are true?

a) The Pearson correlation coefficient provides a measure of the strength of linear association between two variables.

b) The number of NHS psychiatric beds each year was negatively correlated with the number of involuntary admissions for mental disorders per annum.

c) The significance test for the Pearson correlation coefficient is non-parametric.

d) It can be deduced that the decrease in the number of NHS psychiatric beds was caused by a rise in the number of involuntary admissions for mental disorders each year.

Answers

Statements *a* and *b* are true, while *c* and *d* are false.

The Pearson correlation coefficient measures the strength of linear association between two variables (statement *a* is true)—in the example above, the association between the number of NHS psychiatric beds and the combined number of involuntary admissions to NHS and private facilities under the Mental Health Act 1983 each year in England between 1996 and 2006.

For each year there was a pair of observations: the number of NHS psychiatric beds and the number of involuntary admissions for mental disorders. The relation between the number of NHS psychiatric beds and involuntary admissions each year can be represented by a scatter plot (fig 1). The choice of axes for the variables is irrelevant. Furthermore, the temporal nature of the data is ignored—that is, the pairs of points are plotted regardless of the calendar year when they were collected. A straight line is drawn through the points to provide the most appropriate linear association that exists. Correlation is a measure of how closely the points lie to the straight line—that is, the strength of a linear association. Correlation does not provide a measure of the nature of the linear association—that is, the gradient of the straight line drawn through the points.

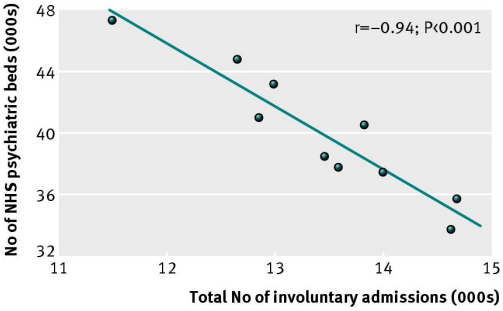

Figure 1 Scatter plot of number of NHS psychiatric beds against combined number involuntary admissions for mental disorders to NHS and private facilities combined per annum. Data were collected for each year between 1996 and 2006

The Pearson correlation coefficient, also known as the product moment correlation coefficient, is represented in a sample by r, while in the population from which the sample was drawn it is represented by ρ. The coefficient is measured on a scale with no units and can take a value from −1 through 0 to +1. If the sign of the correlation coefficient was positive, then a positive correlation would have existed, indicating that those years with a larger number of NHS psychiatric beds were associated with a larger number of involuntary admissions for mental disorders. If the sign of the correlation coefficient was negative, then a

p.sedgwick@sgul.ac.uk

negative correlation would have existed, indicating that those years with a smaller number of NHS psychiatric beds each year were associated with a larger number of involuntary admissions, or vice versa.

If all the points on the scatter plot lay on a straight line, then a perfect correlation would have existed (a correlation coefficient of 1 or −1). A correlation coefficient of zero would have indicated no linear association between the two variables—that is, they are uncorrelated. Figure 2 shows some schematic representations of correlation.

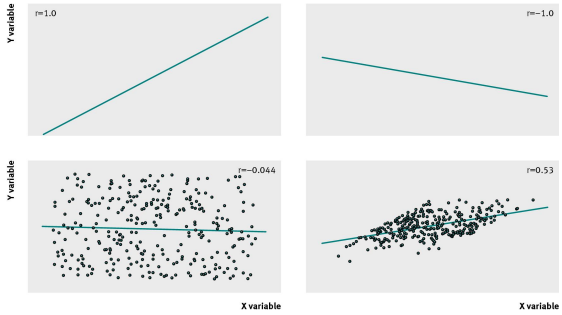

Fig 2 Schematic scatter plots illustrating different values of *r*

A negative correlation existed between the number of psychiatric NHS beds and combined number of involuntary admissions for mental disorders to NHS and private facilities each year (r=−0.94) (*b* is true). The points appear to lie close to the line, suggesting a strong linear association, supported by the correlation coefficient close to −1. As the number of beds decreased over successive years, the number of involuntary admissions rose each year.

A significance test can be undertaken to derive a P value for the correlation coefficient, with statistical hypothesis testing similar to the traditional approach described in a previous Statistical Question.[2] The null hypothesis states that the population correlation coefficient from which the sample was taken is zero. The alternative hypothesis states that the population correlation coefficient from which the sample was taken is not equal to zero; the alternative hypothesis is two sided, and therefore the correlation coefficient may be <0 or >0. The significance test is parametric (*c* is false) and requires at least one of the two variables to be distributed normally. Parametric tests have been described in a previous question.[3] The P value for the significance test of the correlation coefficient (r=−0.94) was <0.001, therefore there was a significant negative correlation between the number of psychiatric NHS beds and the number of combined involuntary admissions for mental disorders to NHS and private facilities each year.

A significant correlation does not mean that a cause and effect relation can be implied (*d* is false). That is, it cannot be inferred that the decrease in the number of NHS psychiatric beds each year was caused by a rise in the number of combined involuntary admissions to NHS and private facilities. Indeed, it would seem illogical to suggest that a decrease in the number of NHS psychiatric beds was caused by an increase in the number of involuntary admissions. The purpose of the study was to examine the number of involuntary admissions under the Mental Health Act 1983 in England and whether the number of NHS psychiatric beds was sufficient to meet these healthcare needs.

In the example above, the correlation coefficient was derived from pairs of measurements: the number of psychiatric NHS beds and the number of involuntary admissions for mental disorders for each year from 1996 to 2006. Each measurement represented the number across England. Correlation coefficients can be derived to describe the linear association between two variables, with pairs of measurements obtained from each person in a sample. However, a correlation coefficient should be derived only if the pairs of measurements are independent. In the example above it is not obvious whether independence existed between pairs of measurements, because presumably patients could have been admitted involuntarily more than once in a year or possibly in more than one year.

The statistical significance of *r* is directly related to the sample size. Small samples require the correlation coefficient to have a larger value—closer to −1 or to 1—for the linear association to be significant. Conversely, in large samples a small value of *r*—closer to zero—may be significant, despite the linear association being weak. This highlights a common misconception regarding correlation coefficients, that a significant correlation coefficient implies that a strong linear association exists between two variables. It is important to review the result of the significance test along with the value of the correlation coefficient and inspection of the scatter plot of the two variables when investigating the strength of a linear association.

Spearman's rank correlation coefficient is a non-parametric equivalent to Pearson's correlation coefficient. Pearson's is calculated if the two variables are numerical and at least one is distributed normally. Spearman's rank correlation coefficient would be calculated if neither variable was distributed normally or if one of the variables was discrete (such as the number of teeth extracted) or was measured on an ordinal scale (such as an anxiety rating score). Spearman's rank correlation coefficient has similar properties to Pearson's correlation coefficient and will be discussed in a future question.

Competing interests: None declared.

1 Keown P, Mercer G, Scott J. Retrospective analysis of hospital episode statistics, involuntary admissions under the Mental Health Act 1983, and number of psychiatric beds in England 1996-2006. *BMJ* 2008;337:a1837.
2 Sedgwick P. Statistical hypothesis testing. *BMJ* 2010;340:c2059.
3 Sedgwick P. Parametric v non-parametric statistical tests. *BMJ* 2012;344:e1753.

Cite this as: *BMJ* 2012;345:e4483