Anais da Academia Brasileira de Ciências (2012) 84(4): 1147-1156
(Annals of the Brazilian Academy of Sciences)
Printed version ISSN 0001-3765 / Online version ISSN 1678-2690
www.scielo.br/aabc

Use of the correlation coefficient in agricultural sciences: problems, pitfalls and how to deal with them

MARCIN KOZAK[1], WOJTEK KRZANOWSKI[2] and MAŁGORZATA TARTANUS[3]

[1]Department of Botany, Warsaw University of Life Sciences - SGGW, Nowoursynowska 159, 20-776 Warsaw, Poland
[2]College of Engineering, Mathematics and Physical Sciences, University of Exeter, EX4 4QF, UK
[3]Institute of Horticulture, Konstytucji 3 Maja 1/3, 96-100 Skierniewice, Poland

Manuscript received on May 30, 2012; accepted for publication on July 10, 2012

ABSTRACT

This paper discusses a number of aspects concerning the analysis, interpretation and reporting of correlations in agricultural sciences. Various problems that one might encounter with these aspects are identified, and suggestions of how to overcome these problems are proposed. Some of the examples presented show how mistaken and even misleading the interpretation of correlation can be when one ignores simple rules of analysis.

Key words: experimental design, linearity, path analysis, Simpson's paradox, statistical analysis.

INTRODUCTION

Pearson's correlation coefficient (hereafter called simply "correlation coefficient") is a statistical method of quantifying the association, or "coherence", between two variables. It is therefore a very popular tool for analysing data that arise in many scientific disciplines, for instance biology (e.g., Soares et al. 2011, Juliá and Peris 2010, Camargo et al. 2011), biomedical and medical sciences (e.g., Camacho et al. 2010, Oliboni et al. 2011), earth sciences (e.g., Fontana et al. 2010, Rangel et al. 2011), social sciences (e.g., Conte et al. 2011), economics (e.g., Misztal 2011), ergonomics (e.g., Bin and Richardson 2010, Zadry et al. 2011), and of course agriculture (e.g., Chełkowski et al. 2000, Cheng et al. 2010, Lakhesar et al. 2010, Bandehagh and Hossein Zadeh Moghbeli 2011, Cherati et al. 2011, Heidari Zooleh et al. 2011, Herrero et al. 2011, Kesavacharyulu et al. 2011, Rogiers et al. 2011).

Because of this pervasive usefulness, the methodology is taught at basic levels and to a very wide audience, even in secondary schools (Holmes 2001). In fact it is so common, so frequently used (if not overused), and so "well-known" as a technique that quite probably many of those who apply it do so rather automatically, without the proper careful consideration of what is being done and how the results should be interpreted. Moreover, various recommendations and rules for interpretation that are offered by some textbooks and webpages may not be particularly accurate or helpful, and these can end up by doing more harm than good.

The purpose of this paper is therefore to assist agricultural researchers in understanding the correlation coefficient, by pointing out some

Correspondence to: Marcin Kozak
E-mail: nyggus@gmail.com

common problems and pitfalls in its use, and by clarifying some characteristics relating to its inference and its interpretation. We will show that sometimes too much credence is given to correlation and its testing, and that proper interpretation is essential if strange and non sensical results are to be avoided (even if one follows commonly advocated rules). We start by briefly elucidating the nature of correlation, and then go on to consider each of the main aspects that we wish to highlight.

What is Correlation?

Whenever a researcher has measured two variables, it is natural to ask whether or not they exhibit any "coherence" – in other words, whether they behave consistently in rising and falling in value together or whether there is no discernible pattern to their joint behaviour. At the theoretical level, such co-association is quantified by the covariance between two random variables. This is the expected value of the product of the deviation from the mean of each variable. If the variables rise together then the two deviations tend to be either positive together or negative together, so their products tend to be positive and the expected value yields a positive covariance. If one rises as the other falls then positive deviations in one variable tend to be associated with negative ones in the other so the products tend to be negative and the expected value yields a negative covariance. On the other hand, if there is no discernible pattern in the behaviour of the two variables then positive and negative products will tend to cancel and the covariance will be near zero. However, the size of the covariance is not predictable in any given situation, other than that its absolute value must be no greater than the product of the standard deviations of the two variables. So dividing the covariance by this product of standard deviations yields the correlation, whose value must therefore lie between -1 and +1.

Theoretically, the correlation coefficient should therefore be limited to describing the association between two variables that are not in a cause-and-effect relationship. In practice, however, it is common to use correlation simply as a measure of strength and direction of a relationship, whether or not it is a cause-and-effect relationship. Examples of when it is such a relationship can be found in numerous papers in which correlation is reported as a part of regression analysis (e.g., Torrico and Janssens 2010, Keiser 2010), but this in general is not how correlation should be applied for the simple reason that it measures different phenomena and between different types of variables than does regression analysis (e.g., Kozak 2008a).

Reasons for this confusion may be found in the calculations that result when the theoretical expectations of random variables are replaced by the averaging of observed variable values. The reasoning outlined above behind theoretical covariance and correlation values will in general only hold good for linear association when it is extended to the averaging of observed variable values. The calculation of the correlation coefficient will only therefore be a meaningful measure of strength of association if the relationship between the two variables is a linear one. This heavy reliance on linearity has linked correlation and regression in researchers' minds, leading to some of the potential problems mentioned above. Indeed, what to do when there is lack of linearity is the first consideration in our main following section.

Problems and Pitfalls with Inference, Interpretation and Presentation of Correlation

Lack of Linearity

If a relationship is non-linear then it may be possible to linearise it, by transforming one or both variables; the most common among such transformations are the logarithmic or root-square ones. Alternatively, if the relationship is not linear but monotonic then Spearman's rank

correlation coefficient is an appropriate measure of association. If neither of these cases pertains, then applying correlation to a non-linear relation will not represent the association between two variables as it is an inappropriate measure to use. This can be well illustrated by Anscombe's (1973) four data sets that are shown in Figure 1. In each panel the value of the correlation coefficient is the same and equals 0.82, being significant at $p = 0.002$. However, the assumption of linear association is valid only for the data in top left panel, while the other data sets violate this basic assumption, either representing non-linear association (top right panel) or being strongly affected or influenced by an outlier (both bottom panels).

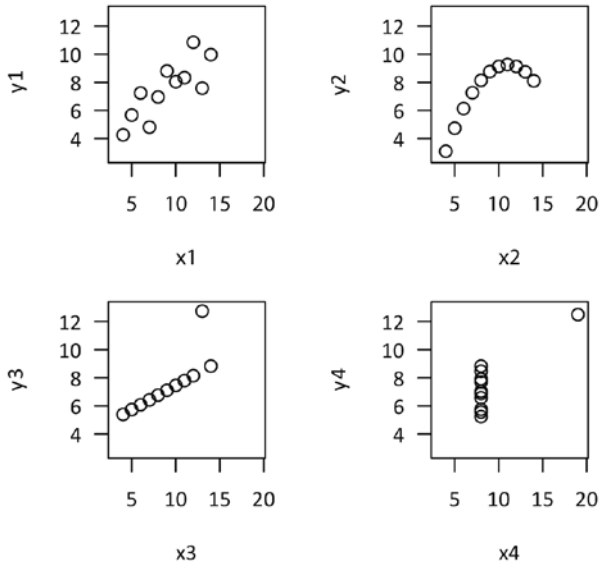

Figure 1 - Anscombe's (1973) four data sets.

Thus, one should always check on linearity of an association in non-massive data before calculating a correlation coefficient, and the best way of doing so is by graphing pairs of variables in a scatterplot (as in Figure 1). By this means one can detect not only the possibility of a non-linear association, but also outlier observations, grouped data and other problems that are described in the following sections.

Kozak and Wnuk (2011) proposed a graphical method of inspecting linearity of associations in multidimensional data sets that uses the fact that for linear association Pearson's and Spearman's correlation should give very similar values. So, a high difference between these two types of correlation coefficient points to a lack of linearity of the association of the two corresponding variables. This method can be particularly useful for data sets with many variables, but it can also be very well seen in Anscombe's quartet; referring to Figure 1, we obtain the following pairs of correlation coefficients:

top left panel	: Pearson = 0.82	Spearman = 0.82	diff = 0.00
top right panel	: Pearson = 0.82	Spearman = 0.69	diff = 0.13
bottom left panel	: Pearson = 0.82	Spearman = 0.99	diff = -0.17
bottom right panel	: Pearson = 0.82	Spearman = 0.50	diff = 0.32

As previously mentioned, therefore, only the data in the top-left panel are appropriate for Pearson's correlation coefficient, and whether the other ones are appropriate for Spearman's correlation requires only checking if the corresponding relationships are monotonic.

Outliers

A single outlier can strongly affect the correlation value. This can be easily seen from Figure 1, where the two bottom panels suffer from outliers.

How outliers can affect correlation analysis has been discussed for example by Kozak (2009c), who argues that merely removing outliers is seldom the best choice (except, of course, when the outlier originates from an error in data entry, but this must be carefully checked). The point is to discover "what the outlier means and what implications it may have for the phenomenon being studied". In fact, outliers can be a source of interesting information on the phenomena being studied, often constituting the most interesting observations in the data set. Not checking data for outlier values can lead to gross mistakes in data analysis and most of all interpretation. We need to emphasize that for

correlation there can be three types of outliers that can affect interpretation: an outlier in the value of the first variable (which is a value that is not typical for this variable), an outlier in the value of the second variable, and an outlier in the bi-variate sense (which is a value that does not have to be atypical for these two variable considered separately, but is atypical for the relationship between them). An example of such a bi-variate outlier can be found in Figure 2.

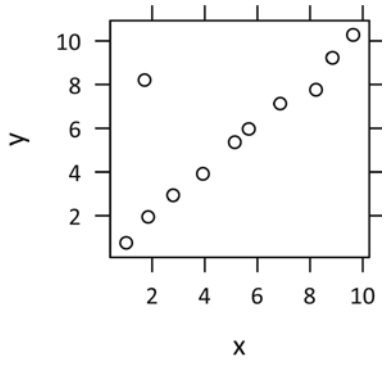

Figure 2 - A bivariate outlier: a value that is not atypical for the distributions of variables *x* and *y* considered separately, but is atypical for their bivariate distribution.

Grouped Data

In agricultural data analysis we often encounter grouped data, the groups representing different populations (e.g., species, factor levels, vegetation seasons, locations etc.). If we operate on raw data from such groups, then before pooling such data (i.e., combining them into one correlation analysis by ignoring the grouping structure), the data must be carefully checked to make sure that there are no major discrepancies between the various within-group associations. For this, the best method is to simply graph the data in a scatterplot or a set of scatterplots.

A simple illustration comes from the Iris data set (Anderson 1935, Fisher 1936). Figure 3 highlights the problem, showing that for the two Iris species *I. setosa* and *I. virginica* the correlation is quite similar (slightly above 0.3, as shown by the practically parallel corresponding lines in the graphs), but for *I. versicolor* it is much higher and reaches almost 0.8. However, when we pool the data, the correlation coefficient increases up to 0.96, suggesting a very strong linear relationship. The reason for this strange-at-first-glance result is the shift of values for *I. setosa* considerably below the other ones. This is a completely technical "trick" that inflates the correlation coefficient to a non sensically high level, and if one does pool the data for correlation analysis, as in the left panel of Figure 3, one makes a methodological mistake. Unfortunately this is quite a common mistake in the applied sciences, agricultural sciences not being an exception.

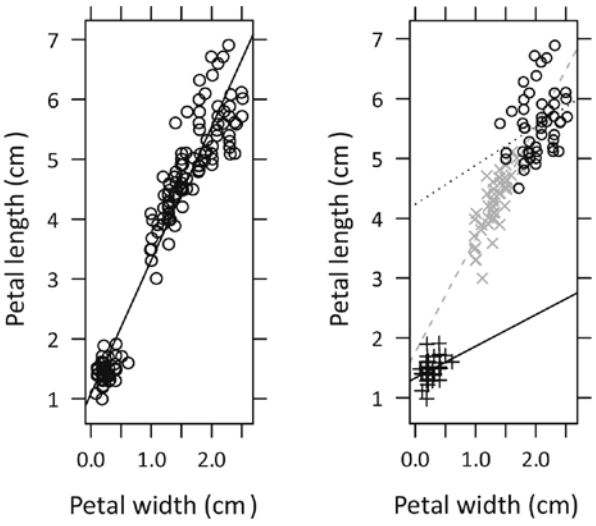

Figure 3 - Association between petal length and width for three Iris species: *I. setosa* (+), *I. versicolor* (×) and *I. virginica* (○). In the left panel, groups (Iris species) are ignored and the correlation coefficient (0.96) is determined for the pooled data. In the right panel, groups are correctly taken into account and correlation coefficient is determined for each species separately, giving 0.33 for *I. setosa*, 0.79 for *I. versicolor* and 0.32 for *I. virginica*.

Such problems in data analysis are often referred to as the Simpson paradox (Simpson 1951). Originally dealing with contingency tables, nowadays the Simpson paradox is also related to associations in general, such as those in Figure 3. In fact, one can often obtain much greater differences between the correlation for pooled data and separate within-group correlation coefficients.

In agricultural research, one must be especially careful when thinking of using correlation for data from designed experiments, because as such they are always grouped (by factor(s) levels). In fact, to make it sensible to employ correlation for designed experiments, ideally one should calculate the coefficient separately for each factor level (for one-way experiments) or factor combination (for more complex experiments). This use of correlation is natural since each factor level represents some population of interest, but the fact that each factor level or combination should be represented by a sufficient number of replications is a problem. A possible alternative is to calculate the coefficient from means of factor levels or factor combinations. However, this use is not so natural because then the assumption for correlation that each observation comes from the same population is violated. Nevertheless, the coefficient can be employed as a summary statistic but care is needed when interpreting it. Moreover, in this case the sample size may be quite small, which presents additional problems for interpretation.

Sample Size

A small sample causes problems for any statistical inference, because estimation will not be precise. For agricultural researchers, this means that their conclusions may be of little value. The case of correlation is no different in general, but it is a particularly critical case because low precision in its estimation can easily lead to rather substantially false conclusions. Small samples can be very dangerous for the correlation coefficient (Kozak 2009a, 2011). It is not uncommon to estimate a correlation coefficient based on 5 or so sample elements, but in this case there is an appreciable probability of obtaining an estimate very far from the true population coefficient.

We will support Kozak (2009a, 2011) results by showing how imprecise confidence interval estimation can be in small samples. According to Kozak (2008b), the most interesting information about correlation (providing that the assumptions have been approximately met) can be obtained from confidence interval estimation. This is because the standard estimator of Pearson's correlation has a non-symmetrical distribution, and so its standard error can be misleading too. A confidence interval for the correlation coefficient, on the other hand, shows how precise the estimation is: it shows the most probable range of the population coefficient, given the confidence level.

Figure 4 shows widths of confidence intervals for the correlation coefficient for different sample correlations and sample sizes. Note that the maximal width of such an interval is 2 (because correlation ranges from -1 to 1). Note also that an interval of width 1 is rather wide: this would mean for example obtaining an interval of (-0.5, 0.5), not really a precise estimate. The graphs show that the width can vary greatly depending on the estimate of the correlation coefficient and on the sample size. The conclusion that can be drawn is clear cut: if one estimates the correlation coefficient from a small sample, one must be aware that a confidence interval for the population correlation will be very wide, meaning imprecise estimation and inference, and so interpretation of little precision. This should always be kept in mind because researchers seldom seem to be aware how imprecise the correlation coefficient is when estimated from a small (and even medium, of size 20-30) sample.

The problem is that it is rather uncommon to estimate confidence intervals for correlation coefficients, no matter how useful is this way of inference for correlation coefficient (Kozak 2008b). Indeed, estimating such an interval, one can see that the estimation is often imprecise; otherwise, if one only reports a correlation coefficient and its corresponding p-value (see the section below), this important information about precision of estimation is usually lost.

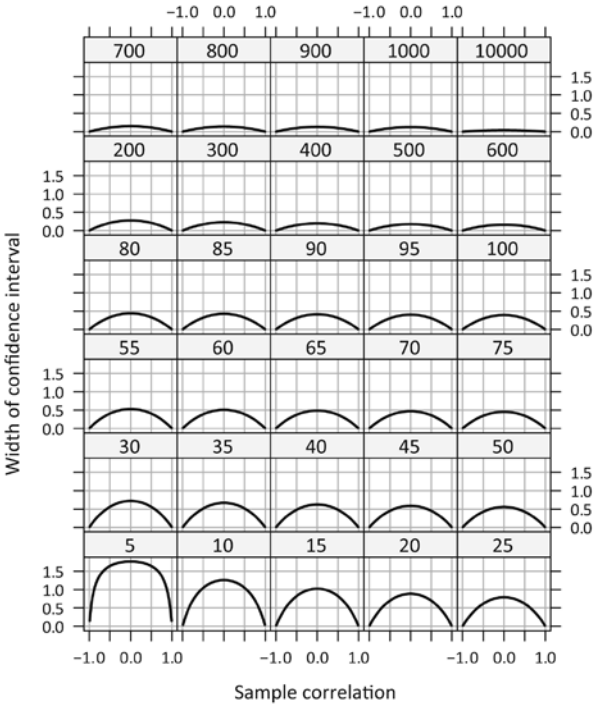

Figure 4 - Width of a 95% confidence interval for the correlation coefficient for samples of sizes of 5 to 10,000. Note that 99% confidence intervals would be even wider.

Hypothesis Testing

Reporting correlations with accompanying results of hypothesis testing is a standard in agricultural research. This is usually done either in the text, by reporting a correlation coefficient with the corresponding p-value for the null hypothesis that the population correlation is zero, or by reporting a correlation table with asterisks to represent statistical significance of the coefficients. Asterisks inform whether a corresponding hypothesis has or has not been rejected, so treat verification of the hypothesis as a black and white situation: the relationship is or is not significant. Kozak (2010) claims that asterisks used in such a context can do more harm than good for the reasons given above – such an emphatic statement (significant / non-significant) disregards sample size and those situations in which the significance is at the boundary level (e.g., p-value is around 0.05). For these reasons Kozak (2010) suggests – wherever possible – giving up asterisks and reporting p-values instead. Confidence intervals are an even better option, especially because p-values are themselves random variables that can be very variable (Cumming 2008).

The problem of hypothesis testing for the correlation coefficient has been implicitly touched on above, while discussing sample size. The general and most common way of viewing hypothesis testing of correlation is to treat the null hypothesis that the population correlation is zero as an indication of lack of linear relationship between the two variables. This is correct, but do note that if the null hypothesis is not true and the population correlation is, say, 0.10, then it does not mean that there is a linear relationship. Kozak (2008b) showed that for very large samples (10,000 in his example), very small sample correlation coefficients can be statistically significant (0.023 in his example). Such significance merely suggests presence of a non-zero population correlation coefficient, not necessarily an "important" or substantial one. On the other hand, for small samples the sample correlation coefficient must be close to 1 to be "significant" ("significant" here meaning statistical significance, which is far from what would be considered significant by our logic – see, e.g., Reese 2004). Thus, hypothesis testing for correlation should always be treated with caution, confidence intervals – as discussed earlier – being a much better option.

Incorrect or Clumsy Interpretation

Correlation is far too often interpreted by applying strict boundaries for its values, for example |r| higher than 0.7 represents very strong association while |r| smaller than 0.3 means very weak or non-existing correlation. Such interpretation is incorrect because it totally disregards the context of the phenomenon studied (Kozak 2009b). A low value of the correlation coefficient can represent incredibly strong association, while a

high value can still represent weak correlation. One can never say that without knowledge of the context of the association.

Simply put, if the correlation is expected to be very high (e.g., close to 1) and in the sample it is not, then such association should not be considered strong (even when the value of the correlation coefficient suggests that, reaching e.g. 0.7 or 0.8), and if the correlation is expected to be near zero, then even the value of sample correlation around 0.3 might be considered high.

For example, if two precise methods of measurement of the same parameter give different results with the correlation around 0.8, is it a strong correlation? Or, if two parameters that should not be related are correlated at the level of 0.4, is it a weak correlation? Responses to these questions should be given in the context of the problem as well as of the study, because possible reasons for such non sensical results can be confounding variables, pooled data (as in Figure 3) or an incorrectly designed study.

Presenting Correlations and Correlation Tables

When reporting correlations, one needs to remember that two issues are important: the size of the sample correlation and its precision. Thus when reporting it, one should provide not only its value, but also at least sample size, confidence interval or significance (although the last option is least preferred, especially when it is provided not as a p-value, but as an asterisk to represent bounds of p-values, as discussed above). In addition, one should remember that for correlation it suffices to use two decimal digits, three or more being redundant and representing only negligible variation (Kozak et al. 2011).

When there are many variables to correlate, often correlation coefficients among them are presented in a so-called correlation table or correlation matrix giving the correlations between every pair of variables. Sometimes a lower (White and Watson 2010) (or upper) triangle is given only because the other triangle contains the very same values; often then the first data row and the last column are removed since they add no information, unless the lower and upper triangles report correlations from two different sets of data (e.g., two years). Sometimes one of the triangle of the correlation table reports sample correlations, while the other triangle their p-values (e.g., Alves et al. 2011).

Interpreting Correlation among Many Variables

If there are even just 5 or 6 variables in an observational study, many researchers are confused when trying to disentangle relationships. Unfortunately, the most frequent approach seems to be one in which all the correlations that are significant in the matrix are picked out and discussed individually, without considering whether, for example, the significant correlation between A and B arises as a consequence of the joint association of A and B with one of the other variables. Consequently, associations are interpreted disregarding any causal links that may be present among them. One can use partial correlations to establish such pathways, but this can be a very laborious process in the absence of prior intuition about the variables and a descriptive approach would be better. In fact, partial correlations are seldom encountered in agricultural applications (for an example see Lorentz et al. 2011), the opposite being the case in psychology (e.g., Rosmarin et al. 2011).

In this regard a few relatively simple ideas have been around for many years. In the first one, if one of the eigenvalues of the correlation matrix is zero, then the corresponding eigenvector elements give the coefficients in an exact linear relationship between the standardised variables. So, looking at eigenvectors corresponding to "small" eigenvalues can be very useful in detecting relationships among

all the (standardised) variables. Hills (1969) showed how a correlation coefficient can be converted into a "distance" between the two variables, so either a metric scaling or a cluster analysis will pick out groups of "similar" variables and thereby simplify the picture. Path analysis (Wright 1921) is one of the most common methods for identifying cause-and-effect associations among a set of variables, also in agriculture (actually, Wright invented it for genetics and published in the *Journal of Agricultural Research*, so path analysis has its origin in agricultural sciences). Path analysis has been very popular in various fields, agriculture not being an exception (e.g., Lakshmi and Padma 2011, Maleki et al. 2011). More recently path analysis is considered a part of a more general method, structural equation modeling (Shipley 2002), with a new estimation methodology and more application possibilities; applications of such approach to path analysis are also becoming more and more popular in the agricultural sciences (e.g., Dhungana et al. 2007, Kozak et al. 2008). It is worth mentioning that criticism of path analysis is practically as old as the method itself (Niles 1923), and does not seem to stop. Some say it is a method of statistical fantasy rather than reasoning (Everitt and Dunn 1991).

CONCLUSION

The correlation coefficient is one of the most often used statistical tools for analysing associations among traits. It is considered simple and intuitive, and it usually is so, but practice shows that far too often it is misinterpreted or misunderstood. But we do believe too that if one is aware of the aspects discussed in this paper, then one should be aware of the traps that exist for the unwary when interpreting correlations.

It is always important to bear in mind any assumptions that underlie the analysis being undertaken or interpretation of any results that have been obtained. We have already stressed the assumption of linear association that is necessary prior to the calculation of the correlation coefficient, and have mentioned the possibility of using Spearman's coefficient for those monotonic relationships that cannot be linearised. However, another important assumption that has not already been mentioned is one that is implicit in any inferential procedure carried out on a sample correlation coefficient. The calculation of the limits in confidence intervals or p-values in hypothesis tests depends on the approximate normality of the Fisher-transformed correlation value (see, e.g., Steel and Torrie 1980, p. 279). The approximation improves as sample size increases, so whereas the inferences for large samples will be reliable, in small samples there may be some inaccuracy and this should be borne in mind.

As a final point, it is worth mentioning a couple of cases where something slightly more complicated than a simple correlation coefficient may be needed. For data sets in which the different pairs of observations are subject either to different precisions or importances, and it is possible to quantify these differences by attaching weights to the observations, then one can calculate a weighted correlation coefficient simply by obtaining the constituent weighted variances and covariances in the usual way. For data sets in which there is no meaningful way of deciding which measurement belongs to which variable, one needs to calculate the intraclass correlation coefficient. A typical example would be when obtaining the correlation between the weights of twins. Here the usual roles of “variables” and “individuals” in correlation are reversed, because we only have one attribute (weight) but two values of it (one on each twin) and there is no meaningful way of saying which twin's weight should be x and which twin's weight should be y. Nevertheless, if we have n pairs of twins, then it is valid to ask what the correlation is for the n pairs of weights. This situation can be thought of as grouped data with two individuals in each of n groups, and is readily extended to the general case with more than two individuals in each group –

e.g. to obtain the association of some genetic trait among all the siblings of n families. Various ways of obtaining such a correlation have been proposed, but nowadays the coefficient is usually estimated using the between-group and within-group mean squares in an analysis of variance (see, e.g., Steel and Torrie 1980, p. 282). However, supplying further details would take us beyond the scope of the present article.

At the very end, it is worth adding that applying correlation does presuppose that the researcher knows what he or she is doing, because if the context for the correlation does not make sense, interpretation of the correlation coefficient will not make sense either.

RESUMO

Este artigo discute uma série de aspectos relacionados a análise, interpretação e forma de relatar correlações em ciências Agrárias. São identificados vários problemas que podem ser encontrados, bem como feitas sugestões de como superá-los. Alguns dos exemplos apresentados mostram quão erradas e mesmo enganosas podem ser as interpretações de correlação quando regras simples de análise são ignoradas.

Palavras-chave: design experimental, linearidade, análise de trilha, paradoxo de Simpson, análise estatística.

REFERENCES

ALVES AA, GUIMARÃES LMS, CHAVES ARM, DAMATTA FM AND ALFENAS AC. 2011. Leaf gas exchange and chlorophyll a fluorescence of *Eucalyptus urophylla* in response to Puccinia psidii infection. Acta Physiol Plant 33(5): 1831-1839.

ANDERSON E. 1935. The Irises of the Gaspé Peninsula. Bull Am Iris Soc 59: 2-5.

ANSCOMBE F. 1973. Graphs in statistical analysis. Am Statistician 27: 17-21.

BANDEHAGH AA AND HOSSEIN ZADEH MOGHBELI AH. 2011. Effects of salinity on wheat genotypes and their genotype × salinity interaction analysis. Res Crops 12(1): 13-19.

BIN SW AND RICHARDSON S. 2010. An ergonomics study of a semiconductors factory in an IDC for improvement in occupational health and safety. Int J Occup Saf Ergon 16(3): 345-356.

CAMACHO S, BERNAL F, ABDO M AND AWAD RA. 2010. Endoscopic and symptoms analysis in Mexican patients with irritable Bowel syndrome, dyspepsia, and gastroesophageal refux disease. An Acad Bras Cienc 82: 953-962.

CAMARGO MGG, SOUZA RM, REYS P AND MORELLATO LPC. 2011. Effects of environmental conditions associated to the cardinal orientation on the reproductive phenology of the cerrado savanna tree *Xylopia aromatica* (Annonaceae). An Acad Bras Cienc 83: 1007-1019.

CHEŁKOWSKI J, KAPTUR P, TOMKOWIAK M, KOSTECKI M, GOLIŃSKI P, PONITKA A, ŚLUSARKIEWICZ-JARZINA A AND BOCIANOWSKI J. 2000. Moniliformin accumulation in kernels of triticale accessions inoculated with F*usarium avenaceum*, in Poland. J Phytopathology 148(7-8): 433-439.

CHENG Z, SALMINEN SO AND GREWAL PS. 2010. Effect of organic fertilisers on the greening quality, shoot and root growth, and shoot nutrient and alkaloid contents of turf-type endophytic tall fescue, *Festuca arundinacea*. Ann Appl Biology 156(1): 25-37.

CHERATI FE, BAHRAMI H AND ASAKEREH A. 2011. Evaluation of traditional, mechanical and chemical weed control methods in rice fields. Aust J Crop Sci 5(8): 1007-1013.

CONTE JC, RUBIO EA, GARCÍA AI AND CANO FJ. 2011. Correspondence model of occupational accidents. An Acad Bras Cienc 83: 1131-1146.

CUMMING G. 2008. Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better. Persp Psychol Sci 3(4): 286-300.

DHUNGANA P, ESKRIDGE KM, BAENZIGER PS, CAMPBELL BT, GILL KS AND DWEIKAT I. 2007. Analysis of genotype-by-environment interaction in wheat using a structural equation model and chromosome substitution lines. Crop Sci 47: 477-484.

EVERITT BS AND DUNN G. 1991. Applied multivariate data analysis. London: Edward Arnold, 464 p.

FISHER RA. 1936. The use of multiple measurements in taxonomic problems. Ann Eugenics 7(2): 179-188.

FONTANA LF, DA SILVA FS, DE FIGUEIREDO NG, BRUM DM, PEREIRA NETTO AD, DE FIGUEIREDO JUNIOR AG AND CRAPEZ MAC. 2010. Superfcial distribution of aromatic compounds and geomicrobiology of sediments from Suruí Mangrove, Guanabara Bay, RJ, Brazil. An Acad Bras Cienc 82: 1013-1030.

HEIDARI ZOOLEH H, JAHANSOOZ MR, YUNUSA I, HOSSEINI SMB, CHAICHI MR AND JAFARI AA. 2011. Effect of alternate irrigation on root-divided Foxtail Millet (Setaria italica). Aust J Crop Sci 5(2): 205-213.

HERRERO N, PÉREZ-SÁNCHEZ R, OLEAGA A AND ZABALGOGEAZCOA I. 2011. Tick pathogenicity, thermal tolerance and virus infection in *Tolypocladium cylindrosporum*. Ann Appl Biology 159(2): 192-201.

HILLS M. 1969. On looking at large correlation matrices. Biometrika 56: 249-253.

HOLMES P. 2001. Correlation: From Picture to Formula. Teach Stat 23(3): 67-71.

JULIÁ JP AND PERIS SJ. 2010. Do precipitation and food affect the reproduction of brown brocket deer *Mazama gouazoubira* (G.Fischer1814) in conditions of semi-captivity? An Acad Bras Cienc 82: 629-635.

KEISER C. 2010. Analysis of Steam Formation and Migration in Firefighters' Protective Clothing Using X-Ray Radiography. Int J Occup Saf Ergon 16(2): 217-229.

KESAVACHARYULU K, REKHA M, BALAKRISHNA R AND SARKAR A. 2011. Association of component characters with leaf yield in advanced generation hybrids of mulberry (*Morus* spp.). Res Crops 12(3): 822-825.

KOZAK M. 2008a. Correlation and regression: similar or different concepts? Stat Transit – new series 9(1): 159-162.

KOZAK M. 2008b. Correlation coefficient and the fallacy of statistical hypothesis testing. Curr Sci 95(9): 1121-1122.

KOZAK M. 2009a. How to show that sample size matters. Teach Stat 31(2): 52-54.

KOZAK M. 2009b. What is strong correlation? Teach Stat 31: 85-86.

KOZAK M. 2009c. Teaching statistics = teaching thinking statistically. Model Assist Stat Appl 4(4): 275-279.

KOZAK M. 2010. Asterisks – friends or foes of statistics? Teach Stat 32(3): 88-89.

KOZAK M. 2011. Online platform supporting teaching correlation. Model Assist Stat Appl 6(1): 71-74.

KOZAK M, AZEVEDO RA, JUPOWICZ-KOZAK J AND KRZANOWSKI W. 2011. Reporting numbers in agriculture and biology: Don't overdo with digits. Aust J Crop Sci 5(13): 1876-1881.

KOZAK M, BOCIANOWSKI J AND RYBIŃSKI W. 2008. Selection of promising genotypes based on path and cluster analyses. J Agric Sci 146: 85-92.

KOZAK M AND WNUK A. 2011. Inspecting associations in multivariate data sets with an interactive modified Bland-Altman plot. Romanian Agric Res 28: 259-262.

LAKHESAR DPS, BACKHOUSE D AND KRISTIANSEN P. 2010. Accounting for periods of wetness in displacement of *Fusarium pseudograminearum* from cereal straw. Ann Appl Biol 157(1): 91-98.

LAKSHMI R AND PADMA V. 2011. Correlation and path analysis studies in chilli in high altitude and tribal zone of Srikakulam district of Andhra Pradesh. Res Crops 12(2): 548-550.

LORENTZ LH, GAYA LG, LUNEDO R, FERRAZ JBS, REZENDE FM AND TÉRCIO MF. 2011. Production and body composition traits of broilers in relation to breast weight evaluated by path analysis. Sci Agric 68(3): 320-325.

MALEKI HH, KARIMZADEH G, DARVISHZADEH R AND SARRAFI A. 2011. Correlation and sequential path analysis of some agronomic traits in tobacco (*Nicotiana tabaccum* L.) to improve dry leaf yield. Aust J Crop Sci 5(12): 1644-1648.

MISZTAL P. 2011. The relationship between savings and economic growth in countries with different level of economic development. e-Finanse 7(2): 17-29.

NILES HE. 1923. The method of path coefficients an answer to wright. Genetics 8: 256-260.

OLIBONI LS, DANI C, FUNCHAL C, HENRIQUES JA AND SALVADOR M. 2011. Hepatoprotective, cardioprotective, and renal-protective effects of organic and conventional grapevine leaf extracts (*Vitis labrusca* var.Bordo) on Wistar rat tissues. An Acad Bras Cienc 83: 1403-1411.

RANGEL KMA, BAPTISTA NETO JA, FONSECA EM, MCALISTER J AND SMITH BJ. 2011. Study of heavy metal concentration and partition in gin the Estrela River: implications for the pollution in Guanabara Bay–SE Brazil. An Acad Bras Cienc 83: 801-815.

REESE RA. 2004. Does significance matter? Significance 1: 39-40.

ROGIERS SY, HOLZAPFEL BP AND SMITH JP. 2011. Sugar accumulation in roots of two grape varieties with contrasting response to water stress. Ann Appl Biol 159: 399-413.

ROSMARIN DH, PIRUTINSKY S, AUERBACH RP, BJÖRGVINSSON T, BIGDA-PEYTON J, ANDERSSON G, PARGAMENT KI AND KRUMREI EJ. 2011. Incorporating spiritual beliefs into a cognitive model of worry. J Clin Psychol 67(7): 691-700.

SHIPLEY B. 2002. Cause and correlation in biology: a user's guide to path analysis, structural equations and causal inference. Cambridge University Press, Cambridge, 317 p.

SIMPSON EH. 1951. The interpretation of interaction in contingency tables. J Royal Stat Soc, Ser. B 13: 238-241.

SOARES V, RODRIGUES FB, VIEIRA MF AND SILVA MS. 2011. Validation of a protocol to evaluate maximal expiratory pressure using a pressure transducer and a signal conditioner. An Acad Bras Cienc 83: 967-971.

STEEL RGD AND TORRIE JH 1980. Principles and procedures of statistics, a biometrical approach. (2nd Edition) New York: McGraw-Hill, 633 p.

TORRICO JC AND JANSSENS MJJ. 2010. Rapid assessment methods of resilience for natural and agricultural systems. An Acad Bras Cienc 82: 1095-1105.

WHITE E AND WATSON S. 2010. An investigation of the relationship between hullability and morphological features in grains of four oat varieties. Ann Appl Biol 156: 281-295.

WRIGHT S. 1921. Correlation and causation. J Agric Res (Wash., D.C) 20: 557-585.

ZADRY HR, DAWAL SZM AND TAHA Z. 2011. The relation between upper limb muscle and brain activity in two precision levels of repetitive light tasks. Int J Occup Saf Ergon 17(4): 373-384.