OPEN ACCESS

ESCOLA DE MEDICINA PUCRS

SCIENTIA MEDICA

Scientia Medica Porto Alegre, v. 31, p. 1-6, jan.-dez. 2021
e-ISSN: 1980-6108 | **ISSN-L:** 1806-5562

http://dx.doi.org/10.15448/1980-6108.2021.1.41065

EDUCATION IN HEALTH SCIENCES

Article numbers as a leading indicator of publication time

Números de artigos como um indicador importante do tempo de publicação

Jimmie Leppink[1]
orcid.org/0000-0002-8713-1374
jleppink@hvvaldecilla.es

Received on: Jun.7th, 2021.
Approved on: Sep. 25th, 2021
Published on: Nov. 5th, 2021.

Abstract

Aims: in health professions education (HPE), the use of statistics is commonly associated with somewhat larger samples, whereas smaller samples or single subjects (i.e., $N = 1$) are usually labelled as needing some kind of 'qualitative' approach. However, statistical methods can be very useful in small samples and for individual subjects as well, especially where we have time series of repeated measurements of the same outcome variable(s) of interest. The aim of this article is twofold: to demonstrate an example of a cross-correlation function for single subjects in a HPE context and to suggest a few settings in HPE where this cross-correlation function can be of use.

Method: the example uses data from a recent Open Access publication on among others article numbers and publication time in a number of major HPE journals to examine the relation between the number of articles published and median publication time over time in the zero-cost Open-Source statistical program R version 4.0.5.

Results: as to be expected, the number of articles published appears somewhat of a leading indicator of publication time: both number of articles in year '*y*' and number of articles in year '*y* minus 1' correlate > 0.6 with median publication time in year '*y*', while correlations of other time differences (e.g., number of articles in year '*y* minus 2' and median publication time in year '*y*', or median publication time in year '*y*' and number of articles in year '*y* plus 1') are substantially smaller.

Conclusion: in line with recent literature, this article demonstrates that the cros-s-correlation function can be used in the context of small samples and single subjects. While the example focusses on article numbers and publication times, it can equally be applied in for example studying relations between knowledge, skills and attitude in individuals, or relations between behaviors of individuals working in pairs or small groups.

Keywords: time series, auto-correlation function, cross-correlation function, leading indicator.

Resumo

Introdução: na educação de profissionais de saúde, o uso de estatísticas é associado comumente a amostras um pouco maiores, enquanto as amostras menores ou assuntos únicos (ou seja, N = 1) são geralmente rotulados como precisando de algum tipo de abordagem "qualitativa". No entanto, os métodos estatísticos podem ser muito úteis em pequenas amostras e para sujeitos individuais, especialmente quando temos séries temporais de medições repetidas da(s) mesma(s) variável(is) de desfecho de interesse. O objetivo deste artigo é demonstrar um exemplo de uma função de correlação cruzada para sujeitos individuais em um contexto de educação de profissionais de saúde e sugerir algumas configurações em que essa função pode ser útil.

Método: o exemplo usa dados de uma publicação recente de acesso aberto sobre, entre outros, números de artigos e tempo de publicação em vários dos principais periódicos da educação de profissionais de saúde para examinar a relação entre o número de artigos publicados e o tempo médio de publicação ao longo do tempo, no programa R versão 4.0.5, programa estatístico de código aberto de custo zero.

Resultados: o número de artigos publicados parece ser um indicador importante do tempo de publicação: tanto o número de artigos no ano "y" quanto o número

[1] University of York, Hull York Medical School, York, North Yorkshire, United Kingdom.

de artigos no ano "y menos 1" se correlacionam > 0,6 com o tempo médio de publicação no ano "y", enquanto as correlações de outras diferenças de tempo são substancialmente menores, como, por exemplo, número de artigos no ano " y menos 2" e tempo médio de publicação no ano " y", ou tempo médio de publicação no ano "y" e número de artigos no ano "y mais 1"').

Conclusão: de acordo com a literatura recente, este artigo demonstra que a função de correlação cruzada pode ser usada no contexto de pequenas amostras e indivíduos únicos. Embora o exemplo se concentre em números de artigos e tempos de publicação, pode igualmente ser aplicado, por exemplo, no estudo de relações entre conhecimento, habilidades e atitudes em indivíduos, ou relações entre comportamentos de indivíduos que trabalham em pares ou pequenos grupos.

ABBREVIATIONS: ACF, auto-correlation function; CCF, cross-correlation function; HPE, health professions education.

Introduction

While statistical methods are widely used in health professions education (HPE), their use in small samples is less common, probably because of a combination of a belief that small samples require 'qualitative' methods and a lack of awareness of statistical possibilities in the face of small samples. Van de Schoot and Miocevic (1) recently edited a very useful book on small sample size solutions, that is available Open Access, and can be used by researchers and practitioners in a wide variety of fields including HPE and the broader education. Especially where we have time series of repeated measurements of the same outcome variable(s) of interest, we have quite a range of methods at our disposal that can help us to study a wide variety of research question and bridge potential gaps between research and practice, even when the sample in a given setting is as small as *one* individual or subject (i.e., $N = 1$).

The aforementioned book provides a range of options for studying effects of interventions as well as for studying relations between variables of interest over time in the case of $N = 1$ or otherwise small samples. One method that is particularly useful to study correlations is the so-called *cross-correlation function* (CCF), which is related to the *auto-correlation function* (ACF) (2). Where we measure the same outcome variable of interest of the same individual or subject on several occasions in time, the results obtained tend to correlate (i.e., *serial correlation*) with results less distant in time (e.g., how we feel today and how we feel tomorrow) usually correlating more strongly than results more distant in time (e.g., how we feel today and how we feel in four days), and this information can be expressed in an ACF of the outcome variable of interest. Next, if on each occasion we measure (at least) two outcome variables of interest, for example A and B, the CCF can help us understand the correlation between A and B and whether substantial changes in A tend to (i.) *precede* changes in B (i.e., A is a *leading* indicator of B), (ii.) occur more or less simultaneously with changes in B (i.e., A and B are contemporary indictors of one another), or (iii.) to be preceded by changes in B (i.e., A is a *lagging* indicator of B, which is the same as saying that B is a leading indicator of A).

The aim of this article is twofold: to demonstrate an example of a CCF of two outcome variables in a HPE context and to provide a few settings in HPE where this CCF can be of use. The example used in this article to demonstrate the CCF comes from an excellent study published Open Access by Maggio and colleagues (3), on among others the number of articles published and publication time in a number of major HPE journals during 2008-2018. Two variables used for the example in this article, the outcomes of which are reported in the first table of the article by Maggio and colleagues, are the total number of articles published in each year in the group of HPE journals included (henceforth referred to as the 'number of articles') and the median publication time in days (henceforth referred to as 'median publication time'). The hypothesis tested is that the number of articles is a *leading* indicator of median publication time, with stronger increases in the number of articles predicting stronger increases in publication time in the year following.

Method

The ACFs of the number of articles and median publication time as well as the CCF of these two outcome variables were computed using the zero-cost Open-Source statistical program

R version 4.0.5 (4) using the following code (no specific package needed to run the code):

```
a <- c(993, 1328, 1578, 1449, 1567, 1796, 1943, 1927, 2126, 2153, 2322)
p <- c(185, 172, 199, 229, 204, 233, 260, 236, 199, 250.5, 251)
acf(a, lag = 5, ylim = range(-1,1))
acf(a, lag = 5, pl=FALSE, test=TRUE)
acf(p, lag = 5, ylim = range(-1,1))
acf(p, lag = 5, pl=FALSE, test=TRUE)
ccf(a, p, lag = 5, ylim = range(-1,1))
ccf(a, p, lag = 5, pl=FALSE, test=TRUE)
```

In this code 'a' and 'p' hold the number of articles (a) and median publication time in days (p) in each year of 2008-2018 as published Open Access by Maggio and colleagues (3) in temporal order reading from left to right, the 'acf' lines are used to obtain ACF plots (the first and third 'acf' line) and coefficients (the second and fourth 'acf' line) for each outcome variable, and the 'ccf' lines are needed to obtain the CCF plot (first 'ccf' line) and coefficients (second 'ccf' line).

Results

Figures 1 and **2** present the ACFs and **figure 3** presents the CCF.

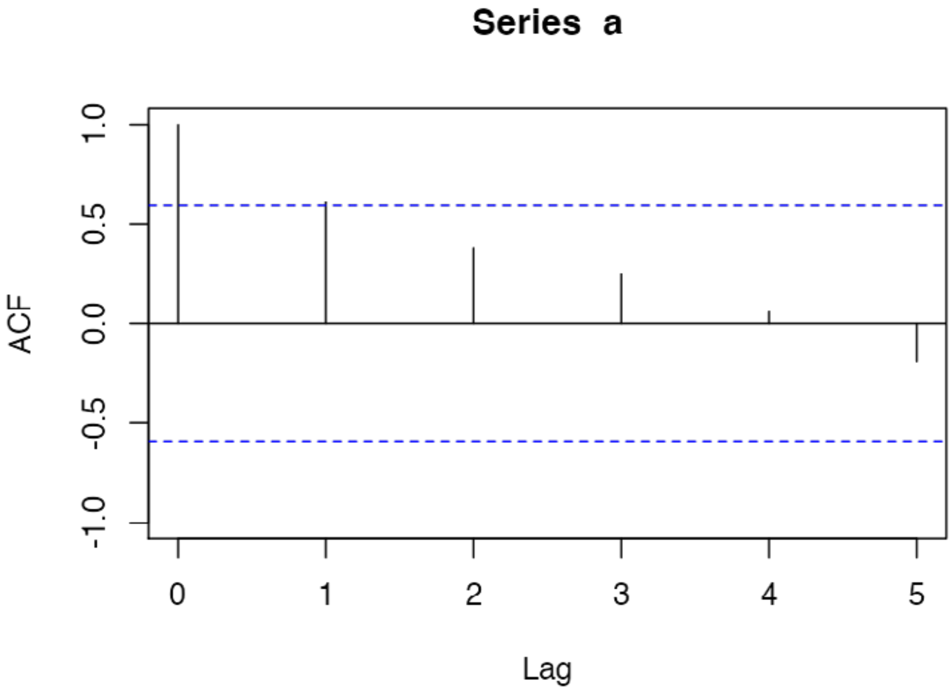

Figure 1 – Auto-correlation function (ACF) articles published (a) and time unit of difference (lag). In this example one lag is one year).

Figure 1 indicates the correlation between the number of articles published in year '*y*' and 1 (i.e., Lag = 1) or several years later, and Figure 2 does the same for median publication time. The first bar in both figures shows a correlation of 1, because that is the correlation with itself. The correlation at Lag 1 in Figure 1 is 0.612, which is statistically significant at the 5% level because it exceeds the blue dotted line that corresponds with a two-sided test at 5%, and the other correlations are smaller which is quite common in time series data: the correlation between the number of articles in year '*y*' and the number of articles in year '*y* plus 1' is stronger than the correlation between the number of articles in year '*y*' and the number of articles more than 1 year away from year '*y*'. In Figure 2, all correlations are within the [-0.4; 0.4] range and not statistically significant at 5%.

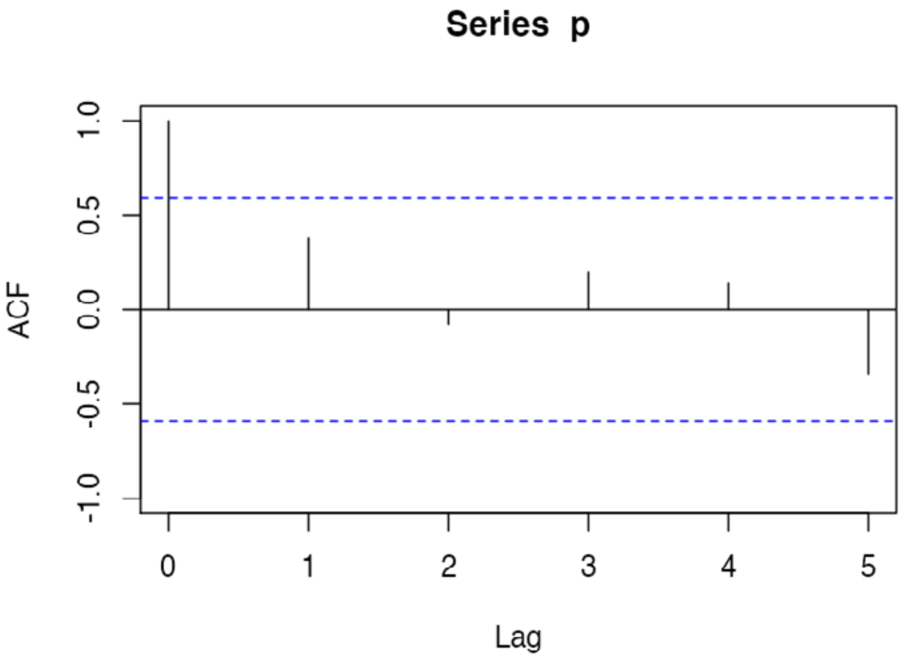

Figure 2 – Auto-correlation function (ACF) median publication time (p) and time unit of difference (lag). In this example one lag is one year.

Figure 3 indicates statistically significant correlations at Lag -1 (0.605) and at Lag 0 (0.711) and not at other lags. Lag 0 expresses the correlation between the outcome variables in year '*y*' (i.e., number of articles as a contemporary indicator of median publication time), whereas Lag 1 expresses the correlation between number of articles in year '*y* minus 1' and median publication time in year '*y*'. As hypothesized, these findings suggest that the number of publications may be somewhat of a leading indicator of median publication time, with stronger increases in the number of articles in one year predicting stronger increases in median publication time in the next year.

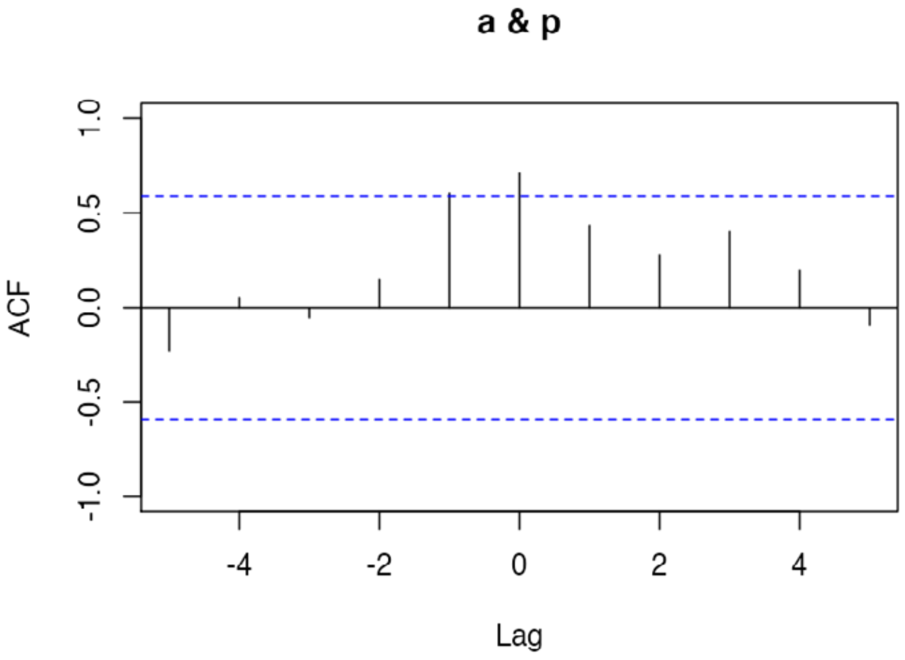

Figure 3 – Cross-correlation function (CCF) articles published (a) and median publication time (p) and time unit of difference (lag). In this example one lag is one year.

Discussion

Maggio and colleagues' important work of documenting numbers and time of publication helps to appreciate how much the volume of articles has grown over the years. As the authors rightly conclude, reasons for the longer publication times remain unknown and deserve further study. While correlation does not equate causation, and other variables will inevitably be needed to understand factors influencing the number of publications and median publication times and how these factors have changed over the years, the CCF can help to appreciate the temporal order of changes in different outcome variables of interest. While the CCF in this case indicates a clear positive correlation at Lag 0 and Lag -1, the other correlations are substantially smaller, indicating among others little relation between median publication time in year '*y*' and number of articles more than one year earlier.

Although the example in this article focusses on publication outcomes recently reported in an Open Access article, the basic idea is that, in line with work published by Van de Schoot and Miocevic (1), the CCF can be used to study correlations where sample sizes are small. In this case, the number of articles and median publication time can be understood as two outcome variables from a single subject of interest, even though that single subject is effectively an aggregate of different journals. At the level of individual learners, CCFs can help among others to study correlations between knowledge, skills and attitudes over time. For example, among medical students, is attitude a leading indicator of knowledge or skill? Is knowledge a leading indicator of skill? Is motivation a leading indicator of performance, is performance a leading indicator of motivation, or do changes in these outcome variables tend to occur more or less simultaneously? Or, in the case the CCF presents strong correlations – positive or negative – at either side of Lag 0, do changes in each outcome variable in some way precede changes in the other outcome variable (e.g., a pattern of increase-decrease waves in both outcome variables but with the waves occurring in turns)?

At a next level, the CCF can also be used to understand relations between behavior of different individuals in a team of two or more people. For instance, in a writing team, do contributions of one author tend to precede contributions by another author? Equally, in settings where questions of possible competition play a role, or of communications in a professional or other type of social network, CCFs can help to understand the temporal order of events over a larger period of time, and as in other settings – including the example used in this article – in combination with other information it may help to shed light on how one event or development may influence change of another kind.

In any case, CCFs are readily available to researchers and practitioners, and provide more information than just a single correlation coefficient alone. For questions where temporal order is of interest, CCFs should be considered as one of the methods to be used.

Notes

Funding

This study was financially supported by the corresponding author.

Conflicts of interest disclosure

The authors declare no competing interests relevant to the content of this study.

Authors' contributions

All the authors declare to have made substantial contributions to the conception, or design, or acquisition, or analysis, or interpretation of data; and drafting the work or revising it critically for important intellectual content; and to approve the version to be published.

Availability of data and responsibility for the results

All the authors declare to have had full access to the available data and they assume full responsibility for the integrity of these results.

References

1. Van de Schoot R, Miocevic M. Small sample size solutions: A guide for applied researchers and practitioners. New York: Routledge; 2020. 285 p. https://doi.org/10.4324/9780429273872

2. Brockwell PJ, Davis RA. Time series: Theory and methods (2nd ed.). New York: Springer Verlag; 1991.

3. Maggio LA, Bynum WE, Schreiber-Gregory DN, Durning SJ, Artino AR. When will I get my paper back? A replication study of publication timelines for health professions education research. Perspect Med Educ. 2020;9:139-46. https://doi.org/10.1007/s40037-020-00576-2

4. R Core Team. R: A language and environment for statistical computing [Internet] Vienna: R Foundation for Statistical Computing (version 4.0.5). Available from: https://www.r-project.org

Jimmie Leppink

PhD in Statistics Education, LLM in Forensics, Criminology and Law, and MSc in Psychology and Law from Maastricht University, the Netherlands; MSc in Statistics from Catholic University of Leuven, Belgium; currently Research Director at Hospital Virtual Valdecilla (HvV), Santander, Spain.

Mailing address

Jimmie Leppink

Hospital virtual Valdecilla

39008

Santander, Spain

Os textos deste artigo foram conferidos pela Poá Comunicação e submetidos para validação do autor antes da publicação.