

Statistical relationship of some basic bibliometric indicators in scientometrics research

Ali Uzun

Middle East Technical University, Department of Statistics
Inonu Bulvari 06531 Ankara, Turkey

Abstract

This paper presents the results of a search for statistical relationships between authorship in terms of the number of authors, number of significant terms/words in article titles, number of references cited, and citation impact of a set of 467 articles published in the international journal *Scientometrics* from 1999 to 2003. Our analysis shows that, controlling for the growth dynamics of citations, the mean citation impact of articles depends strongly on the authorship as measured by the number of authors per article. A chi-square test indicated that the number of references cited in articles and their citation impact are not independent. There is a fairly high degree of linear association between the number of times an article is cited and the number of references it contains. Our analysis also shows that at any level of aggregation there is no statistically significant association between citation impact and the number of significant words in the titles of articles.

1. Introduction

The question of relationship between the number of times a paper is cited and the number of references it contains was first posed in the 1960s [1]. A recent article presented figures reflecting increases in the mean citation impact of papers parallel with the increasing number of references they listed [2]. It is generally hypothesized, but untested, that collaborative research is more visible than sole-authored research as measured by citation impact or other means [3-6]. Much has been written about title lengths/words of journal articles but no research has been carried out in search of any association between titles and citations of articles [7-9]. The motivation for writing this paper is grounded in answering the following three research questions:

- (1) Is there a statistically significant relationship between citation impact and number of authors of articles?
- (2) Is there any relationship between citation impact and the number of references cited in research articles?
- (3) Is there any relationship between citation impact and the numbers of significant words in titles of articles?

2. Research method

Based on the Web of science of Thomson ISI, a total of 467 articles from the journal *Scientometrics* for the five-year time window 1999-2003, have been collected and examined. For each article, the number of authors, number of references it contains, number of significant terms/words, i.e., noun phrases in the titles, and number of times the article has been cited during the interval between the year of publication and the year 2005, have been noted. It should be remarked that the articles published in the year 2003- the most recent articles that we considered- are expected to attain their maximum

number of citations in 2005 since annual citation counts typically peak at around the third year after publication [10].

Recognizing that the statistical relationship is applicable to groups of publications, 467 articles have been classified by the number of authors, cited references, and title words. Mean citation impact of articles within each category have been calculated and adjusted for the growth dynamics from 1999 to 2005. Chi-square tests and the method of least-squares have been employed in search of statistical relationships between citation impact, authorship, cited references and title words of articles. The study has two main limitations. The first is the coverage. It covers only the research articles (ignores reviews, letters, notes, etc.) from a single journal -the international journal *Scientometrics*. The second limitation is the time span. It is limited to the 5-year interval from 1999 to 2003 for the count of articles and 1999 to 2005 for their citations. It is generally agreed that citation statistics produced by time windows shorter than three years may not be sufficiently stable [11]. The time window 1999-2005 is long enough to collect many citations even for the articles published in 2003.

3. Citation impact and references of articles

Table 1 shows a classification of the 467 articles according to the number of references they contain and the number of citations they received during the interval between the year of publication and the year 2005. The expected frequencies are given in parentheses in Table 1 and are obtained by applying the cell-count rule given by

$$\text{Expected cell count} = \frac{(\text{cell row total})(\text{cell column total})}{\text{grand total}}$$

Table 1: Contingency table of observed and expected frequencies of articles, 1999-2005

Number of references	Number of citations				Number of articles
	0	1-2	3-4	5 or more	
0-10	58 (36.4) ^a	36 (42.9)	16 (21.4)	29 (38.3)	139
11-20	38 (42)	50 (49.6)	29 (24.9)	44 (44.5)	161
21-30	14 (22.2)	31 (26.2)	11 (13.1)	29 (23.5)	85
31 or more	12 (21.4)	27 (25.3)	16 (12.6)	27 (22.7)	82
Number of articles	122	144	72	129	467 ^a

^a The figures in the parentheses are the expected cell frequencies

A chi-square test for the independence of row categories (number of references) and column categories (number of citations per year) is performed. The test statistic is given by

$$\chi^2 = \sum \frac{(f_0 - f_e)^2}{f_e}$$

where f_0 denotes the observed frequency and f_e denotes the expected frequency for a particular cell. Large values of χ^2 leads to rejection of the null hypothesis that the number of citations and the number of references of articles is independent. It should be remarked that the chi-square distribution is a good approximation for a count data provided that the count for each cell is at least 5. In our case the

minimum cell count is 11 and thus the cell-count rule is satisfied. The degrees of freedom (d.f.) of χ^2 is $v = (r-1)(c-1) = (4-1)(4-1) = 9$, where r and c are the number of rows and columns of the 4x4 contingency table, respectively. Using the figures in Table 1, the value of the test statistic is found to be $\chi^2 = 24.6$. The critical value for 0.01 level of significance and d.f. of 9 is $\chi^2_{0.01, 9} = 21.7$. So, we conclude that the citation impact of articles as classified in Table 1 is significantly related to the number of references they contain.

Based on the latter conclusion, we assumed an approximate linear relation between the mean number of references per article- the independent variable X- and the number of times it is cited- the dependent variable Y- as given in Table 2. Looking at the pairs of values of X and Y in the table we see that the linearity assumption for the relation between X and Y appears to be reasonable.

Table 2: Classification of references and citations of articles for the period 1999- 2005

Number of references	Number of articles	Mean number of references per article (X)	Mean number of citations per article-year (Y)
0- 10	137	4.0	0.41 ^a
11- 20	166	14.0	0.40
21- 30	86	22.0	0.60
31 and above	78	45.4	0.60
Overall	467	17.8 ^b	0.54 ^b

^a The means in this column are weighted for the number of articles in each reference category and adjusted for the growth dynamics of citations. The adjustment is made using an average annual growth rate of about 0.2 of the Impact Factor of the journal *Scientometrics* from 1999 to 2005.

^b The overall means are also weighted.

We fitted a straight line to the data in Table 2 in the form $Y = A + BX$, where A, B are the least-square estimates calculated using the values of X and Y. The equation that provides the best possible fit is

$$Y = 0.394 + 5.1 \times 10^{-3} X.$$

It should be pointed out that no particular cause-effect pattern between the variables X and Y is implied by this statistical relation. The designations "dependent variable" and "independent variable" carry no connotation that changes in X cause changes in Y. The correlation coefficient which measures the degree of linear association between the variables is found to be $r = 0.799$.

4. *Citation impact and authorship*

To examine whether a statistical relationship exists between the citation impact and number of authors of articles, and what the nature of that relationship may be, the 467 articles is classified by the authorship and citation- count categories. The results of the classification are presented in Table 3.

Table 3: Number of authors and number of citations/article-year aggregated for 1999- 2005

Number of authors per article	Number of articles	Mean number of citations per article per year(adjusted) ¹
1	195	0.46
2	145	0.55
3	89	0.60
4 or more	38	0.67
Overall	467	0.67 ¹

¹See notes a and b to Table 2.

It can be observed by glancing at the figures in the first and last columns of Table 3 that there is a positive linear relation between the two variables- the number of authors per article X- and the mean number of citations/ article-year- Y. To quantify the relationship and its functional form we employed the method of least-squares similar to the one in the previous section. The resulting equation is $Y = 0.400 + 6.8 \times 10^{-2}X$ and $r = 0.994$. The same comments made on the cause-effect relation for the number of references and counts of citations in the previous section go with the authorship and citations.

5. Citation impact and title words

Table 4 shows a two-way classification of the 467 articles with respect to the number of significant words in their titles and the number of citations they received starting with the year of publication till the year 2005.

Table 4: Number of significant words in titles and citation counts of articles, 1999-2005

Number of words	Number of citations				Number of articles
	0	1-3	4-6	7 or more	
1-5	44 (29) ¹	27 (42)	20 (24)	25 (21)	116
6-10	56 (66)	111 (96)	55 (56)	44 (48)	266
11 or more	16 (21)	30 (30)	23 (18)	16 (16)	85
Number of articles	116	168	98	85	467

¹ See note a to Table 1.

A chi-square test for the independence of the word and citation categories is performed using the data in Table 4. The value of the chi- square statistic is computed to be $\chi^2 = 16.0$ whereas the table value with 0.01 level of significance and 6 degrees of freedom is $\chi^2_{0.01,6} = 16.81$. Given these two values and the data in Table 4 it can be concluded that the number of significant title words and citation impact of journal articles are statistically independent.

6. *Concluding remarks*

Our study has revealed that the number of references contained in research articles and their mean citation impact are not independent. An analysis using the method of least-squares has shown that the number of references as independent variable and the citation impact (adjusted) as the dependent variable are linearly associated with a correlation coefficient, $r = 0.799$. Likewise, the number of authors of articles and their citation impact showed a stronger positive correlation with $r = 0.994$. The number of significant words in titles of articles ranged from 2 to 23 and the number of times they were cited ranged from 0 to 50. It is interesting to observe that an article with a title of two words received a total of 50 citations whereas many articles with 15 to 23 significant title words received no citations. A chi-square test applied for the hypothesis of independence of title words and citation counts resulted in acceptance. So, based on the population of 467 articles it can be concluded that the latter variables are not dependent and hence no association is possible between them.

It is worth noting that the statistical associations sought by the analysis presented above provide no more than approximations of the past relationships. Such relationships can be used as basis for predictions, by extrapolation, assuming that the publication and citation practices of authors will remain stable in the future. It should also be noted that the citation impact of journal articles are determined by many factors. The three factors, i.e., authorship, bibliographic references, and title words considered in the present study are deemed to be the most essential. As a future work one may be interested in incorporating more variables through a multivariate regression analysis to explain, in more detail, the observed variances in the citation impact of articles.

References

1. Derek J. De Solla Price, Network of scientific papers, *Science*, 149: 510515, 1965.
2. P. Vinkler, Dynamic changes in the chance for citedness, *Scientometrics*, 54: 421434, Kluwer and Akademiai Kiado, 2002.
3. D. deB. Beaver, Collaboration and team work in physics, *Czechoslovak Journal of Physics*, B 36: 1418, Academia, 1986.
4. M.H. Medoff, Collaboration and the quality of economics research, *Labor Economics*, 10: 597608, 2003.
5. Persson, O., Glanzel, W., Dannel, R. Inflationary bibliometric values: The role of scientific collaboration and the need for relative indicators in evaluative studies, *Scientometrics*, 60: 421432, Kluwer and Akademiai Kiado, 2004.
6. J. Rigby and J. Edler, Peering inside research networks: Some observations on the effect of the intensity of collaboration on the variability of research quality, *Research Policy*, 34: 784794, 2005.
7. G. Lewison and J. Hartly, What's in a title? Numbers of words and the presence of colons, *Scientometrics*, 63: 341356, Kluwer and Akademiai Kiado, 2005.
8. M. Yitzhaki, Relation of the title length of a journal article to the length of the article, *Scientometrics*, 54: 435447, Kluwer and Akademiai Kiado, 2002.
9. B. A. Lipetz, Aspects of JASIS authorship through five decades, *Journal of the American Society for Information Science*, 50: 9941003, John Wiley & Sons, Inc., 1999.
10. J.Adams, Early citation counts correlate with accumulated impact, *Scientometrics*, 63: 567581, Kluwer and Akademiai Kiado, 2005.
11. E. Erkut, Measuring Canadian Business School Research Output and Impact, *Canadian Journal of Administrative Sciences*, 19: 97123, Dollhouse University, 2002.