Adjusted Age Distribution and Its Application to Impact Factor and Immediacy Index

Isao Asai
Department of Industrial Engineering, University of Osaka Prefecture, Mosu, Sakai, Osaka, 591, Japan

For the age count between citing and cited papers, the difference in the year of publication is widely used. In some cases, the frequency obtained is inaccurate because of age bias. This article describes a probabilistic method for adjusting the raw frequency, and its application to the impact factor and immediacy index in journal evaluation. The main procedure is that 1/8 of the raw frequency at any given age is assigned to the previous age and to the following age, respectively. Adjusting the frequency over all ages of use, the curve representing the age distribution becomes smoother.

Introduction

The age distribution of references and citations represents the pattern of use of papers in a given topic over time. For the age count between citing and cited papers, the difference in the year of publication is widely used. This measurement is simple and easy to use. However, the frequency obtained is inaccurate because of age bias.

Investigating the period count based on the difference in the month of publication, such age bias can be illustrated. Consider, for example, three papers A, B, and C published in January 1979, December 1979, and January 1980, respectively. In the case of A and B, the age count is zero although the period count is 11. In B and C, the age count is one although the period count is one. This fact suggests that the age count based on the year has an irrational aspect.

Generally speaking, the period count based on the month produces more accurate statistics than that based on the year; similarly, the date produces more accurate statistics than the month. Investigation of the month or date of publication in all citing and cited papers, however, is practically impossible. Therefore, it is necessary to eliminate the age bias without recollecting the detailed data.

The purpose of this article is to develop a method for adjusting the raw frequency using probability theory and to demonstrate its application to the impact factor and immediacy index in journal evaluation.

Received November 30, 1979; revised May 22, 1980; accepted July 10, 1980

A Probabilistic Approach to Age Bias

A probabilistic approach is effective in eliminating the age bias. Consider the difference in the month of publication of two papers as a random variable. If we assume that the probability that a paper is published in any given month is 1/12, and that the months of publication of two such papers are independent of each other, then the number of all possible month pairs is 144.

There are two papers linked together by reference or citation; the citing paper is published in month m_1 in year n_1, and the cited paper is published in month m_2 in year n_2. The probability $P(X)$ of the random variable $X(= m_1 - m_2)$ in age count $t(= n_1 - n_2)$ is given as follows:

$$
\begin{aligned}
P(X=-11) &= 1/144 \quad (1,12),\\
P(X=-10) &= 2/144 \quad (1,11),(2,12),\\
P(X=-9) &= 3/144 \quad (1,10),(2,11),(3,12),\\
&\vdots\\
P(X=-1) &= 11/144 \quad (1,1),(2,2),(3,3),(4,4),\\
P(X=0) &= 12/144 \quad (5,5),(6,6),(7,7),(8,8),\\
P(X=1) &= 11/144 \quad (9,9),(10,10),(11,11),(12,12),\\
&\vdots\\
P(X=11) &= 1/144 \quad (12,1),
\end{aligned}
$$

where (m_1,m_2) represents a month pair of m_1th and m_2th month.

These probabilities show that the probability density distribution of period count in a given age approximates an isosceles triangle, as shown in Figure 1. If the same type of analysis is made of the date of publication, the shape will be even clearer.

Figure 1 shows the place and the proportion in which the age bias appears. It is reasonable to assume that the range of the random variable in the age count t is $-6 < X \leqslant 6$. The inconsistency of age count occurs at both ends to the left and to the right of the distribution. The

 CCC 0002-8231/81/030172-03$01.00

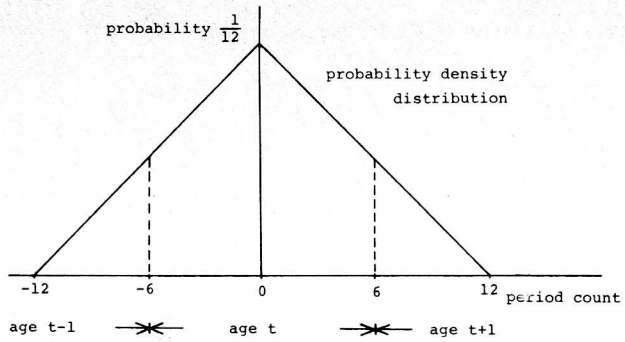

FIGURE 1. Probability density distribution of period count in age t.

events in which $X \leqslant -6$ belong essentially to the previous age count $t-1$, and the proportion obtained from the area of a triangle in the left end is 1/8 of the area of the isosceles triangle. On the other hand, the events in which $X > 6$ belong to the age count $t+1$, and the proportion, the area in the right end, is 1/8.

These analyses suggest, in a probabilistic sense, that 1/8 of the frequency of use counted at any given age has to be assigned to the previous age and to the following age, respectively.

Adjusted Age Distribution

The raw distribution is adjusted over all ages of use. Note that in the case where the age count t is zero, the period of use is never negative. This means that $X \geqslant 0$ always. The probability density distribution in that case is represented by the right side of the isosceles triangle in Figure 1. Therefore, the raw frequency y_t of use at any given age t is altered to the adjusted frequency $\bar{y}_t$ by the following equations:

TABLE 1. Raw and adjusted frequencies of citations to the *Journal of the American Society for Information Science* in 1978.

age t	raw frequency y_t	adjusted frequency $\bar{y}_t$
0	11	13.6
1	43	41.3
2	50	48.3
3	43	42.5
4	32	34.6
5	42	38.6
6	25	27.3
7	26	24.4
8	14	13.9
9	1	2.5
10	0	0.1

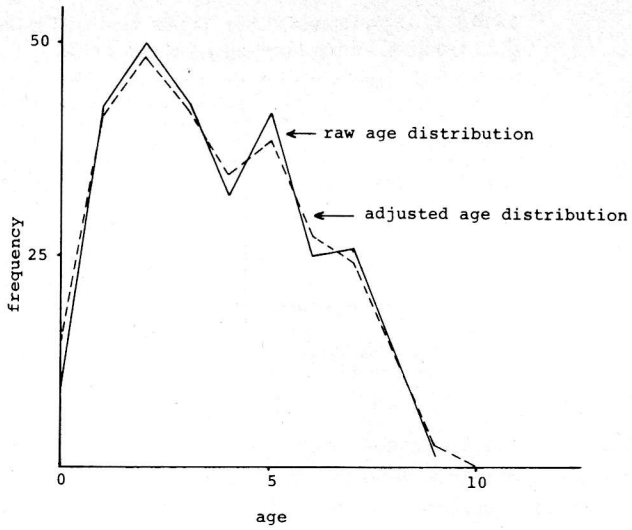

FIGURE 2. Raw and adjusted age distributions of citations to the *Journal of the American Society for Information Science* in 1978.

$$\bar{y}_t = \tfrac{1}{8} y_{t-1} + \tfrac{6}{8} y_t + \tfrac{1}{8} y_{t+1}, \quad 1 < t < c+1,$$

$$\bar{y}_0 = \tfrac{3}{4} y_0 + \tfrac{1}{8} y_1, \quad t = 0,$$

$$\bar{y}_1 = \tfrac{1}{4} y_0 + \tfrac{6}{8} y_1 + \tfrac{1}{8} y_2, \quad t = 1,$$

$$\bar{y}_{c+1} = \tfrac{1}{8} y_c, \quad t = c+1,$$

where c is the maximum value of age. The adjusted age distribution can be obtained by plotting the adjusted frequency $\bar{y}_t$ instead of the raw frequency y_t.

An example is taken from *Journal Citation Reports* (JCR), 1978 annual volumes of *Social Sciences Citation Index* (SSCI) [1]. Table 1 gives the raw and adjusted frequencies of citations to the *Journal of the American Society for Information Science.* The adjusted frequency in the age of four is calculated as follows:

$$(43 + 6 \times 32 + 42)/8 = 34.625.$$

Figure 2 shows the raw and adjusted age distributions. Adjusting the frequency over all ages of use, the curve representing the age distribution becomes smoother.

Application to Impact Factor and Immediacy Index

This section attempts to define two measures which are conceptually equivalent to Garfield's impact factor and immediacy index, and to demonstrate an experiment for making clear the differences between these measures.

In journal evaluation, Garfield's measures are described by the formulas:

$$\text{impact factor} = \frac{f_{k-1} + f_{k-2}}{u_{k-1} + u_{k-2}},$$

$$\text{immediacy index} = f_k/u_k,$$

TABLE 2. An example of the impact factor (IF), the adjusted impact factor (AIF), the immediacy index (II), and the adjusted immediacy index (AII) in 1978.

	citations in 1978				source items			measures in 1978			
journal title	78	77	76	75	78	77	76	IF	AIF	II	AII
HARVARD LAW REV	132	485	382	335	44	41	32	11.877	11.418	3.000	3.628
PSYCHOL REV	29	180	182	253	35	20	33	6.830	6.710	0.829	1.264
YALE LAW J	112	335	131	206	53	37	36	6.384	6.322	2.113	2.375
ADV EXP SOC PSYCHOL	5	37	48	56	7	10	6	5.313	5.164	0.714	1.196
COGNITIVE PSCHOL	15	100	104	205	18	18	23	4.976	5.070	0.833	1.319

where f_k is the frequency of citations in year k and u_k is the number of source items in year k [2].

Garfield's measures use the raw frequency of journal citations based on the year. Because the year in this case equals the age, the frequency needs adjusting to eliminate the age bias. The two adjusted measures for use in journal evaluation can be expressed by the following equations:

$$\text{adjusted impact factor} = (\tfrac{1}{4} y_0 + \tfrac{7}{8} y_1 + \tfrac{7}{8} y_2 + \tfrac{1}{8} y_3)/ (s_1 + s_2)$$

$$\text{adjusted immediacy index} = (\tfrac{3}{4} y_0 + \tfrac{1}{8} y_1)/s_0,$$

where y_t is the frequency of citations in age t and s_t is the number of source items in age t. These are adjusted measures of the average number of times a paper in a given journal is cited during a particular age.

The experiment showing the differences between these measures uses the data of 1212 journals in common with the JCRs 1977 and 1978 annual volumes of SSCI [1]. Twelve data sets are collected from the JCRs, including citations in 1977 to items from 1974 to 1977, citations in 1978 to items from 1975 to 1978, and source items published from 1975 to 1978. Eight measures are obtained from these data sets, including the impact factor (IF), the adjusted impact factor (AIF), the immediacy index (II), and the adjusted immediacy index (AII), in 1977 and 1978, respectively. Table 2 presents a list of five journal titles involving four measures in 1978 and seven related data sets.

To investigate the relationships among these eight measures, the correlation coefficient is computed as shown in Table 3. The correlation between IF and AIF in Table 3 is 0.999 in 1977 and 0.998 and 1978. Because of this very high correlation, it might not make much difference in the results whether the raw or adjusted age distribution is adopted.

However, there are two important differences between II and AII. One difference appears in the relationship between the IIs in 1977 and 1978. The correlation varies from 0.593 to 0.710. This means that the relationship between the AIIs in 1977 and 1978 is stronger than that between the IIs in 1977 and 1978. The other difference appears in the relationship between the IF and II in the same year. The correlation varies from 0.670 to 0.801 in 1977 and from 0.642 to 0.779 in 1978. This suggests that the relationship between the AIF and AII is stronger than that between the IF and II. Because of these differences, it is suggested that the decision of whether to adopt the raw or adjusted age distribution may have an effect on the accuracy of the results.

TABLE 3. Correlation matrix showing relationships among eight measures: impact factor (IF), adjusted impact factor (AIF), immediacy index (II), and adjusted immediacy index (AII), in 1977 and 1978, respectively.

		1977			1978			
		AIF	II	AII	IF	AIF	II	AII
1977	IF	0.999	0.670	0.789	0.861	0.868	0.699	0.803
	AIF		0.687	0.801	0.865	0.872	0.695	0.797
	II			0.969	0.560	0.569	0.593	0.641
	AII				0.659	0.668	0.646	0.710
1978	IF					0.998	0.642	0.778
	AIF						0.652	0.779
	II							0.917

Summary

In order to eliminate the age bias generated by the irrational aspect of the age count based on the year of publication, probability theory may be used. One-eighth of the raw frequency at any given age is assigned to the previous age and to the following age, respectively. The adjusted age distribution obtained from this procedure smooths the raw age distribution. It is suggested that the use of the adjusted age distribution may make a difference in the accuracy of results in the case of relationships between the original and adjusted immediacy indexes.

References

1. *Social Sciences Citation Index, Journal Citation Reports, 1977 and 1978 Annual Volumes.* Philadelphia: Institute for Scientific Information; 1977–1978.
2. Garfield, E. "Citation Analysis as a Tool in Journal Evaluation." *Science.* 178:471–479; 1972.

Adjusted Age Distribution and Its Application to Impact Factor and Immediacy Index

Isao Asai

Reprinted from
JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE
Vol. 32, No. 3, May 1981, pp. 172–174