A Note on the Bias in Herfindahl-type Measures Based on Count Data

Bronwyn H. Hall[1]
University of California at Berkeley and NBER
September 2000 (revised July 2001, January 2005)

Abstract

A Herfindahl index of constructed from shares based on count data where the number of counts is small will generally be biased downward because of the statistical properties of count data and Jensen's inequality. This note suggests a simple correction for the bias and illustrates its applicability when using measures based on patent data and patent citation data.

Keywords: patent data, citation data, herfindahl, generality, count data

1. Introduction

Measures based on citations obtained by patents in individual patent classes or held by individual firms often suffer from bias due to the count nature of the underlying data. The source of the bias is the fact that cells with small numbers of expected citations have a non-zero probability that no citations will actually be observed. When this happens, the cell is removed from the analysis, implying that measures of diversification will be biased downward and measures of concentration will be biased upwards. For example, the widely used patent generality or originality measures defined in Henderson, Jaffe, and Trajtenberg (1998) take the form of diversification measures and will therefore be biased downward when the total number of citations to or from the patent are small. If the bias is not corrected for, patents with few forward or backward citations will be more likely to be considered less "general" or "original" than those with many.

This note suggests a method for correcting the bias that is valid under a set of simple but fairly general assumptions. The two key assumptions are the following:

[1]An earlier version of this note was published as an appendix to Chapter 13 of Jaffe and Trajtenberg (2001). Note that the data provided with that book and on the NBER website at http://www.nber.org/patents contain measures that correct for the bias discussed in this note.

1. Either we treat the total number of citations (or patents) on which the measure is based as given (that is, we condition on them) or the number is large enough relative to the individual cell counts so that it can be treated as non-random.
2. The probability that a given citation or patent falls in a cell is independent of the probability that it falls in another cell. That is, there is no causal connection between the deviation of the observed outcome from the expected outcome in a particular cell and what happens in another cell (other than the adding up constraint). We can therefore describe the probability distribution over a set of cells as a set of multinomial probabilities.

Given these assumptions, I am able to compute a simple correction for the bias that depends only on the total number of counts in the measure. This correction is large when the number of counts is small and quickly converges to zero as the number of counts increases.

Mathematically, the statement of the problem is the following: suppose a researcher uses a Herfindahl-type measure to describe the concentration of patents or cites across patent classes, patent holders, or some other set. Here I use patents as an example, but all the same arguments apply to citation counts. For a set of N patents falling into J classes, with N_j patents in each class ($N_j \geq 0,\ j=1,...,J$), the sample Herfindahl index (HHI) for diversification across the classes is defined by the following expression:

$$HHI = \sum_{j=1}^{J} \left(\frac{N_j}{N} \right)^2 \tag{1}$$

However, the population Herfindahl for patents with this value of diversification is given by

$$\eta = \sum_{j=1}^{J} \lambda_j^2 \tag{2}$$

where the λ_js are the multinomial probabilities that the N patents will be classified in each of the J classes. Under reasonable assumptions, $E[N_j/N]=\lambda_j$. Unfortunately, this does NOT imply that $E[HHI|N]=\eta$ because of nonlinearity. In fact, in general the measured HHI will be biased upward when N is small, due to Jensen's inequality and the properties of the count distribution.

2. Computing and adjusting for the bias

Assume a multinomial distribution with parameters (λ_j, $j=1,...,J$) for the $\{N_j\}$; then the expectation for each N_j^2 is the following (see Johnson and Kotz 1969):

$$E[N_j^2 \mid N] = N\lambda_j + N(N-1)\lambda_j^2 \tag{3}$$

Conditional on the total number of patents N, this implies the following relation between the estimated and true Herfindahl measure:[2]

$$\begin{aligned} E[HHI \mid N] &= E\left[\sum_{j=1}^{J}\left(\frac{N_j}{N}\right)^2 \Big| N\right] = \sum_{j=1}^{J}\frac{E\left[N_j^2 \mid N\right]}{N^2} = \sum_{j=1}^{J}\frac{N\lambda_j + N(N-1)\lambda_j^2}{N^2} \\ &= \frac{1}{N} + \frac{N-1}{N}\sum_{j=1}^{J}\lambda_j^2 = \frac{1}{N} + \frac{N-1}{N}\eta \end{aligned} \tag{4}$$

Note that as $N \uparrow \infty$, $E[HHI|N] \rightarrow \eta$, as we would expect. The bias in this estimator of η is the following:

$$E[HHI \mid N] - \eta = \frac{1-\eta}{N} \tag{5}$$

The bias declines at a rate N as the number of counts grows and as concentration increases. Both results are intuitive.

Under the assumptions given in the introduction, it is straightforward to correct for this bias. Consider the following estimator for the Herfindahl:

$$\hat{\eta} = \frac{N \cdot HHI - 1}{N-1} \tag{6}$$

For a given N, and under the assumption that the underlying process is multinomial with parameters (λ_j, $j=1,...,J$) this estimator is an unbiased estimator of η:

$$E[\hat{\eta} \mid N] = \frac{N \cdot E[HHI \mid N] - 1}{N-1} = \frac{1 + (N-1)\eta - 1}{N-1} = \eta \tag{7}$$

It is also true that standard error estimates based on these measures and obtained in the conventional way will be biased, but it is also possible to compute the exact relationship between the standard error estimated from biased measures and that estimated for the unbiased measures. The standard error of the estimated mean of the Herfindahl will

[2] Conditioning on N is innocuous unless the process that generates the total number of draws (patents or citations) is related to the particular set of multinomial parameters with which we are working. For example, the procedure outlined here may not be valid if "general" patents (patents whose cites are widely distributed across patent classes) are also highly cited patents. I am grateful to Tom Rothenberg for a discussion of this point.

be biased downward by *(N-1)/N*. This is large if *N* is small and it does not depend on the estimated Herfindahl. An unbiased estimator for the variance of the mean Herfindahl over a set of *M* observations is the following:

$$\widehat{Var}[\bar{\eta}] = \frac{1}{M}\sum_{k=1}^{M}\frac{N_k^2 \cdot Var(HHI_k)}{(N_k - 1)^2} \tag{8}$$

where HHI_k is the *k*th biased estimate of the Herfindahl. Of course, if one uses the unbiased estimator to form the mean, one does not need to perform this correction in addition.

3. The generality index

For many purposes, the measure used is one minus the Herfindahl rather than the Herfindahl itself. For example, Hall, Jaffe, and Trajtenberg (2001) and Henderson, Jaffe, and Trajtenberg (1998) define generality as

$$G_i = 1 - \sum_{j=1}^{J}\left(\frac{N_{ij}}{N_i}\right)^2 \tag{9}$$

where N_i denotes the number of forward citations to a patent, and N_{ij} is the number received from patents in class *j*. Patents with a high value of G_i are cited across a broad range of patent classes.

This measure is also a biased estimate of the true measure $\gamma_i = 1 - \eta_i$:

$$E[G_i \mid N_i] = 1 - E\left[\sum_{j=1}^{J}\left(\frac{N_{ij}}{N_i}\right)^2 \Big| N_i\right] = 1 - \frac{1 + (N_i - 1)\eta_i}{N_i} = \frac{N_i - 1}{N_i}\gamma_i \tag{10}$$

The bias is the following:

$$E[G_i \mid N_i] - \gamma_i = -\frac{\gamma_i}{N_i} \tag{11}$$

Again, the absolute size of the bias declines as the sample size increases, and as generality decreases. The generality index will be biased downward in general and this effect is larger for small *N*. Figure 1 shows a 3-dimensional plot of the bias versus the index and values of N. Clearly the magnitude is largest when *N* is small or generality is high.

It is straightforward to compute an unbiased estimator of γ_i:

$$\hat{\gamma}_i = \frac{N_i}{N_i - 1}G_i \tag{12}$$

As in the case of the Herfindahl, the true standard errors of mean generality indices will be *N/(N-1)* larger than the estimated standard errors. When the number of cites to a patent is small, generality will be underestimated and it will be more likely that significant differences among generalities of different patents will be found. But as already indicated, correcting for the bias is straightforward.

4. Examples

I conclude this paper with a couple of illustrations that show the magnitude of the bias arising from using an uncorrected version of the generality index. In the first example, I simply illustrate the aggregate effect of the bias over time as the number of citations per patent changes. In the second, which is drawn from Mowery and Ziedonis (2002), I show how the existence of the bias can impact conclusions drawn from small samples of data based on patents with varying numbers of citations.[3]

The first example uses all the US patent data for patents applied for between 1975 and 1997 that were granted between 1975 and 2002, drawn from Hall, Jaffe, and Trajtenberg (2001) and updated to 2002. Figure 2 shows a graph of average generality and bias-corrected generality for all the patents that are cited more than once between their application date and 2002.[4] The figure also shows the average number of cites received by these patents on the right hand scale. The cites peak at nearly 11 per patent in 1986, and then decline, slowly at first, and then more rapidly after about 1994, due to the truncation at 2002.

Because the average number of citations per patent is falling over time, the bias in the generality measure will get worse over time. This is visible in the figure, which shows average uncorrected generality declining rapidly after about 1992. The bias-adjusted generality also declines, but only for patents applied for in the last couple of years, for whom only about half the eventual expected citations have been observed. Also, because of application-grant lags and the drop in the number of citations per patent, the data in 1996 and 1997 are based on only about a third to a half of the number of patents in the

---

[3] I am very grateful to David Mowery and Arvids Ziedonis for making the raw data on which their paper is based available to me.

[4] Although generality can theoretically be computed when there is only one citation (it will be exactly zero), the bias correction is not defined, and the quantity itself is not very meaningful.

earlier years. The main lesson from this figure is that we should be cautious about drawing conclusions about declines in generality over time in the presence of truncation, even if we are able to do a first-order bias correction.

The second example uses data that does not suffer from serious truncation bias, but where the conclusions are based on relatively small numbers of patents. Mowery and Ziedonis (2002) studied the impact of the Bayh-Dole Act in the United States on patenting by three major U.S. research universities: the University of California, Stanford University, and Columbia University.[5] Using the bias-adjusted generality index, they looked at changes in the generality of patents granted to inventors at these universities before and after the Act, relative to the generality of a set of control patents. The top panel of Table 1 of this paper reproduces Table 6 of their paper, which shows a series of t-statistics for tests of the hypothesis that the generality of university patents from various periods is different from that of the controls.

The table shows that after the Bayh-Dole Act, the generality of patents issued to all three universities was higher than that of the control patents, and that generality increased for UC and Stanford patents. However, the result seems to be true primarily for non-biomedical patents, whereas biomedical patents show insignificant increased generality, and their generality is not much greater than that of the control patents (and in fact, in the case of Stanford, is somewhat less in some cases).

The middle panel of the table shows the results of the Mowery and Ziedonis test statistics computed using a generality index that has not been adjusted for bias, and the bottom panel shows the number of patents on which each test is based (sample plus controls). As expected, more of the tests are significant in the middle panel (6 tests gain a * and only one loses a *) and those most affected are those based on smaller number of patents. Although the main conclusions of the paper would not have been changed by using the measure that is not corrected for bias, several results for the biomedical patents have changed significantly: in particular, we might have concluded that the UC biomedical patents had become more general and those from Stanford less general, whereas the bias-corrected results suggest a much more moderate change, if any.

---

[5] The reader is referred to their paper and the book by Mowery *et al* (2004) for details on this study.

References

Hall, Bronwyn H., Adam Jaffe, and Manuel Trajtenberg. 2001. "The NBER Patent-Citations Data File: Lessons, Insights, and Methodological Tools," in Jaffe and Trajtenberg, *Patents, Citations, and Innovations*, Cambridge, MA: The MIT Press. Data available at http://www.nber.org/patent and http://emlab.berkeley.edu/users/bhhall/data.html

Henderson, Rebecca, Adam Jaffe, and Manuel Trajtenberg. 1998. 'Universities as a Source of Commercial Technology: A Detailed Analysis of University Patenting 1965-1988," *Review of Economics and Statistics* 80: 119-127.

Johnson, Norman L., and Samuel Kotz. 1969. *Discrete Distributions*, New York: John Wiley and Sons.

Mowery, David C., Richard R. Nelson, Bhaven Sampat, and Arvids A. Ziedonis. 2004. *Ivory Tower and Industrial Innovation: University-Industry Technology Transfer Before and After the Bayh-Dole Act*, Stanford, CA: Stanford University Press.

Mowery, David C. and Arvids A. Ziedonis. 2002. "Academic Patent Quality and Quantity before and after the Bayh-Dole Act in the United States," *Research Policy* 31(3): 399-418.

Figure 1

Bias of the Generality Index Based on Count Data

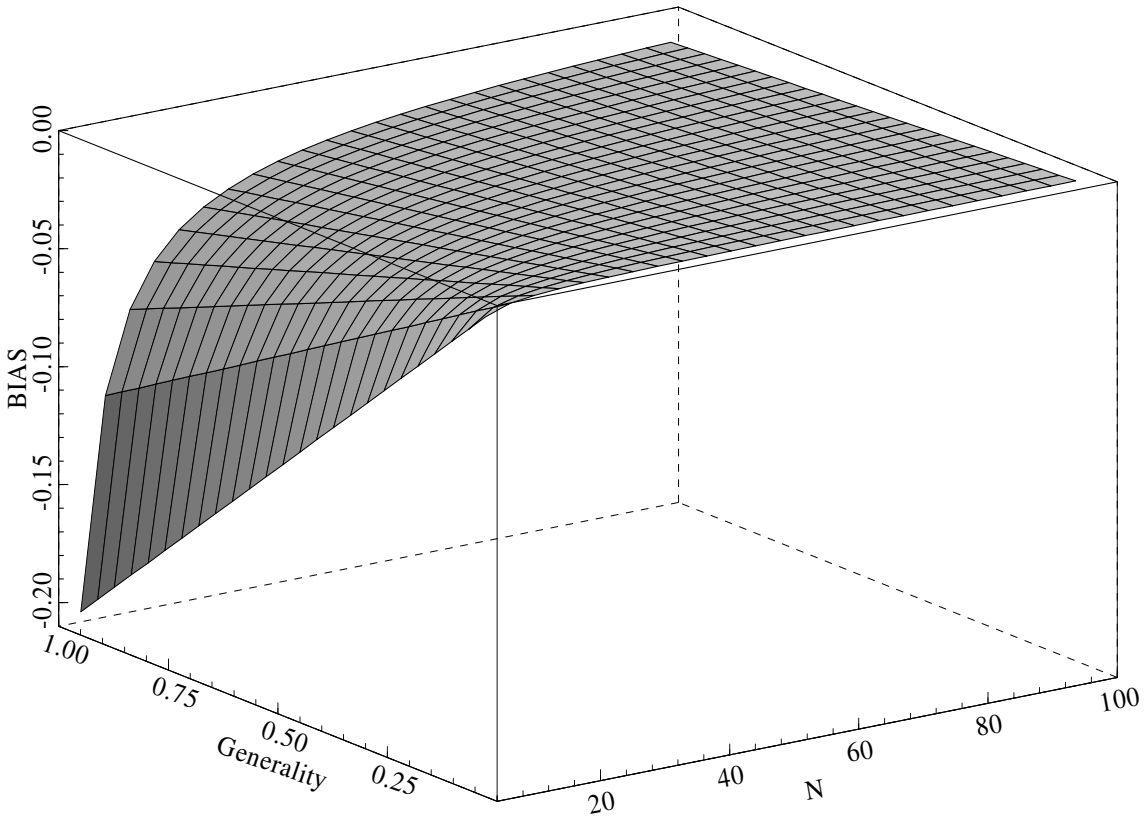

Figure 2

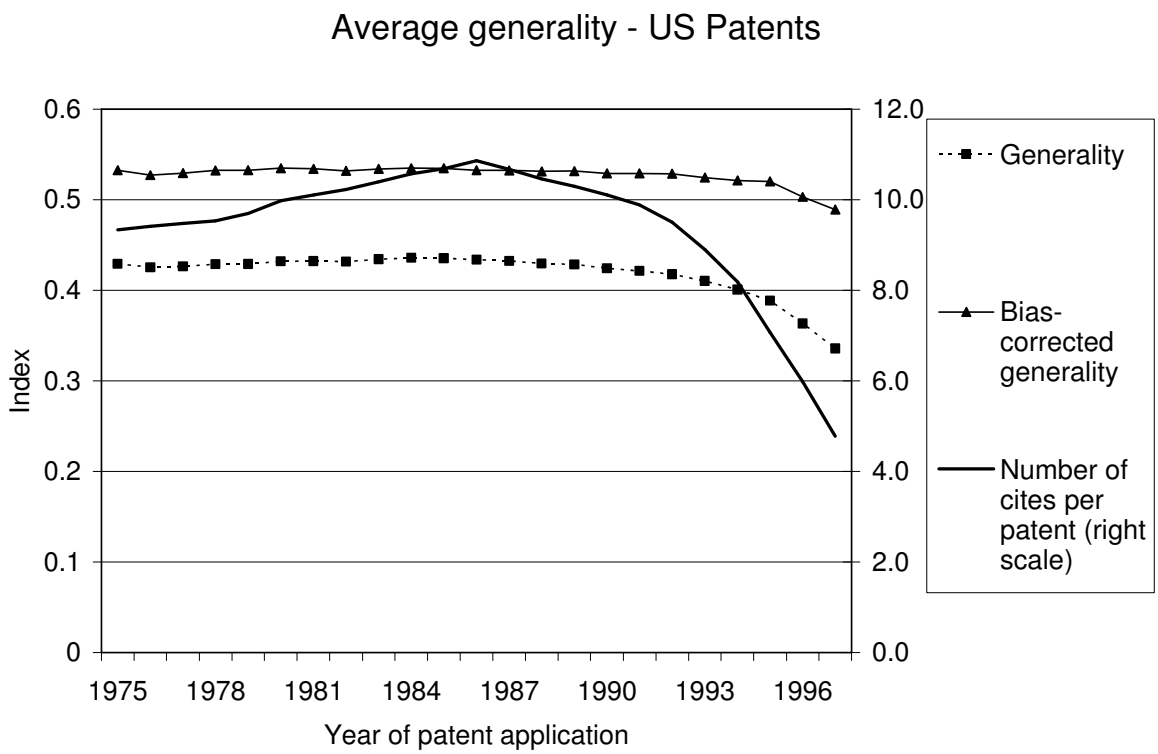
Average generality - US Patents
0.6
0.5
0.4
0.3
0.2
0.1
0
Index
12.0
10.0
8.0
6.0
4.0
2.0
0.0
1975
1978
1981
1984
1987
1990
1993
1996
Year of patent application
Generality
Bias-corrected generality
Number of cites per patent (right scale)

Table 1

Table 6 from Mowery and Ziedonis (2002)

	Overall			Biomedical			Non-Biomedical		
	UC	Stanford	Columbia	UC	Stanford	Columbia	UC	Stanford	Columbia
Patent application 1970-1980; patent issued 1975-1980	-0.35	2.31**		0.19	1.56		-0.70	1.80*	
Disclosed before 1981; patent application 1981-1992	0.08	1.25		-0.32	-1.53		0.91	2.69**	
Disclosed and applied for 1981-1992	3.23**	4.83**	2.49**	1.61	1.76*	1.97**	3.19**	4.58**	1.58

Table 6 using Unadjusted Generality

	Overall			Biomedical			Non-Biomedical		
	UC	Stanford	Columbia	UC	Stanford	Columbia	UC	Stanford	Columbia
Patent application 1970-1980; patent issued 1975-1980	-0.29	3.02**		0.23	1.97*		-0.64	2.41**	
Disclosed before 1981; patent application 1981-1992	1.24	1.87*		0.61	-1.87*		1.65	3.47**	
Disclosed and applied for 1981-1992	3.61**	5.54**	2.69**	1.82*	1.55	2.14**	3.52**	5.53**	1.71*

Sample Size

	Overall			Biomedical			Non-Biomedical		
	UC	Stanford	Columbia	UC	Stanford	Columbia	UC	Stanford	Columbia
Patent application 1970-1980; patent issued 1975-1980	144	167		67	35		75	130	
Disclosed before 1981; patent application 1981-1992	47	50		37	14		8	34	
Disclosed and applied for 1981-1992	401	662	173	240	155	90	159	505	81

*P>.10
**P>.05

Source: top panel from Mowery and Ziedonis (2002); middle and bottom panels are author's computations using the same data. The sample sizes given are the number of patents used in constructing the t-test, including both sample patents and control patents. Therefore they do not include the patents with zero or one citation, for which generality cannot be computed.