

# A Note on the Bias in the Herfindahl Based on Count Data

Bronwyn H. Hall  
UC Berkeley and Nuffield College, Oxford

September 2000

## 1 Introduction

The problem is the following. We would like to use a Herfindahl-type measure to describe the concentration of patents or cites across patent classes, patent holders, or some other set. I will 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 ( $N_j \geq 0, j = 1, \dots, J$ ), the sample Herfindahl index ( $HHI$ ) is defined by the following expression:

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

However, the population Herfindahl is given by

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

where the  $\lambda_j$ s are the multinomial probabilities that the  $N$  patents will be classified in each of the  $J$  classes. Under reasonable assumptions,

$$E \left[ \frac{N_j}{N} \right] = \lambda_j$$

Unfortunately, this does NOT imply that

$$E[HHI] = \eta$$

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 the bias

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

$$E[N_j^2] = N\lambda_j + N(N-1)\lambda_j^2$$

Conditional on the total number of patents  $N$ , this implies the following relation between the estimated and true Herfindahl measure:<sup>1</sup>

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

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

$$E[HHI|N] - \eta = \frac{1 - \eta}{N}$$

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

Figures 1 and 2 show some sample simulations for  $J = 3$  and  $J = 10$ .

## 3 Adjusting for the bias

Consider the following estimator for the Herfindahl:

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

For a given  $N$ , and under the assumption that the underlying process is multinomial with parameters  $\lambda_j, j = 1, \dots, J$ , this estimator is an unbiased estimator of  $\eta$ :

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

---

<sup>1</sup>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.

**Bias of HHI Based on Patent Counts**  
**3 Classes,  $\lambda_2 = \lambda_3 = (1-\lambda_1)/2$**

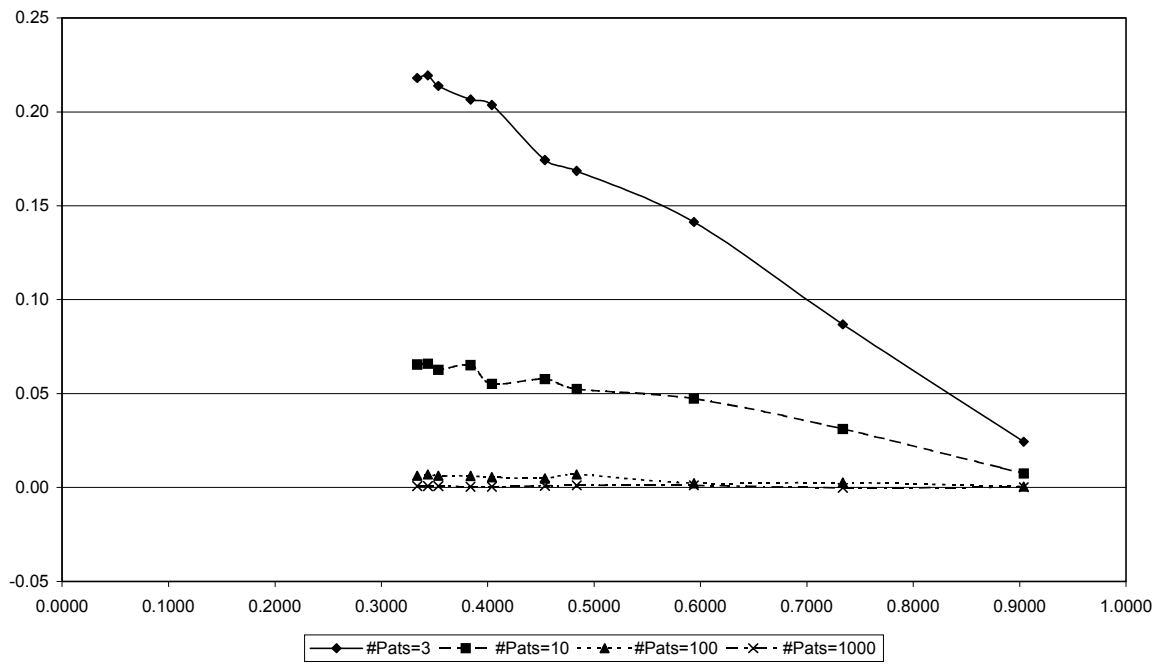


Figure 1: Bias in HHI - 3 Classes

### Bias of HHI Based on Patent Counts 10 Classes

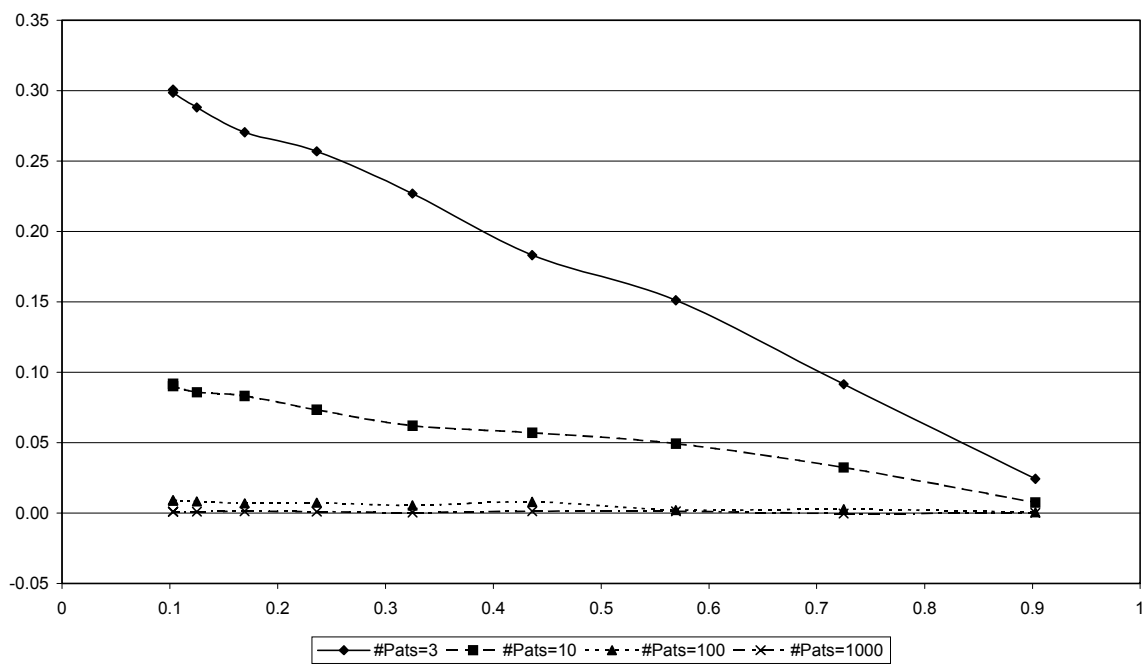


Figure 2: Bias of HHI - 10 classes

## 4 Standard error estimates

Because these estimators are biased, it is also true that standard error estimates obtained in the conventional way will be biased. Figure 3 shows a simulation, 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. Assume we are working with data where each observation on the Herfindahl is based on  $N$  draws and there are  $M$  observations. Then equation (1) tells us that

$$Var(\widehat{\eta}) = \frac{N^2 \cdot Var(HHI)}{(N-1)^2}$$

The implication is that the standard error of the estimated mean of the Herfindahl will be biased downward by  $(N-1)/N$ . This is large if  $N$  is small and 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:

$$Var(\widetilde{\eta}) = \frac{1}{M} \sum_{k=1}^M \frac{N_k^2 \cdot Var(HHI_k)}{(N_k-1)^2}$$

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.

## 5 The generality index

Many researcher use a measure computed as one minus the Herfindahl rather than the Herfindahl itself. For example, Henderson, Jaffe, and Trajtenberg (1998) define generality as

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

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  $\gamma_i = 1 - \eta_i$ :

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

The bias is the following:

$$E[G_i|N_i] - \gamma_i = -\frac{\gamma_i}{N_i}$$

**Bias of Standard Error for HHI Based on Patent Counts**  
**3 Classes,  $\lambda_2 = 0.1 \cdot \lambda_1$**

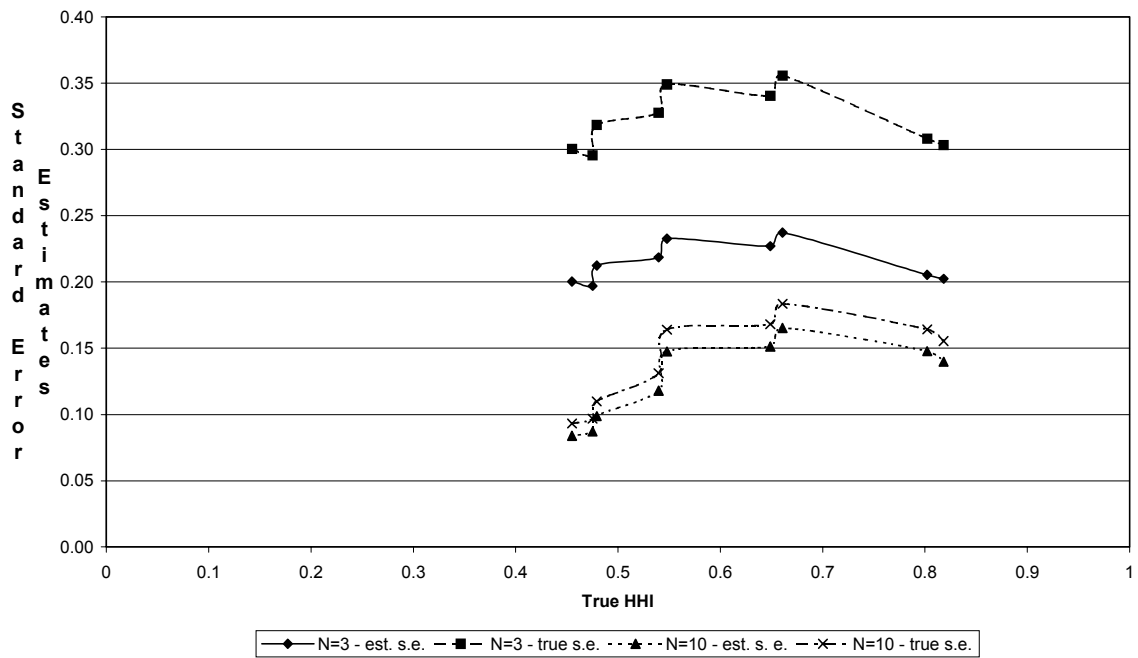


Figure 3: Bias in the standard error estimates

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  $N$ . Once again, one can form an unbiased estimator of  $\gamma_i$ :

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

The same arguments as the previous apply to standard error estimates of the generality index. The true standard errors 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 I have indicated, correcting for the bias is straightforward.

## 6 References

Henderson, Rebecca, Adam B. Jaffe, and Manuel Trajtenberg. 1998. "Universities as a Source of Commercial Technology: A Detailed Analysis of University Patenting, 1965-88." *Review of Economics and Statistics* \_\_:119-127.

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