Inequality and Scientific Productivity

Paul D. Allison

Social Studies of Science, Vol. 10, No. 2 (May, 1980), 163-179.

Stable URL:
http://links.jstor.org/sici?sici=0306-3127%28198005%2910%3A2%3C163%3AIASP%3E2.0.CO%3B2-9

Social Studies of Science is currently published by Sage Publications, Ltd..

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at http://www.jstor.org/about/terms.html. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at http://www.jstor.org/journals/sageltd.html.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is an independent not-for-profit organization dedicated to creating and preserving a digital archive of scholarly journals. For more information regarding JSTOR, please contact jstor-info@jstor.org.

http://www.jstor.org/
Mon Apr 26 10:50:27 2004

● *ABSTRACT*

A new statistic — the modified coefficient of variation — is proposed for measuring the inequality of publication and citation counts. Its advantages include scale invariance, equal sensitivity at all levels of the distribution, allowance for random variation in publication or citation counts, and a close relationship to a probability distribution which fits observed productivity distributions. Also discussed are disciplinary differences in the distribution of productivity and the functional relationship between productivity and its determinants.

Inequality and Scientific Productivity

Paul D. Allison

No matter how it is measured, there is enormous inequality in scientists' research productivity. Derek Price first observed that only about six percent of publishing scientists produce one-half of all papers published.[1] Britton noted that the distribution of publications (as summarized by the Gini index) shows more inequality than the distribution of personal income.[2] Allison and Stewart demonstrated that counts of citations to scientists' work are even more unequally distributed than counts of publications.[3] While such extreme inequality is of interest in itself, several researchers have examined *variations* in inequality in order to test theories about fundamental social processes in science. For example, the closely related notions of the Matthew effect and cumulative advantage suggest that inequality of both publications and citations ought to increase as a cohort of scientists grows older.[4] The Coles and Dietrich, on the other hand, have used the inequality of cita-

Social Studies of Science (SAGE, London and Beverly Hills), Vol. 10 (1980), 163-79

tions to papers and to persons as an indicator of the level of consensus within disciplines and research areas.[5]

In spite of the growing importance of inequality as a concept in the sociology of science, there has been little discussion of the relative merits of different statistical measures of inequality. Faia used the variance to measure inequality of publications.[6] Gaston divided the variance by the mean.[7] Allison and Stewart, the Cole group, and others have used the Gini index.[8] In this paper I consider the appropriateness of these and other measures of inequality for measuring the inequality of scientific productivity. On several grounds, the best measure appears to be a variant of the coefficient of variation which, in its simplest form, is just the standard deviation (not the variance) divided by the mean. In reaching this conclusion, I also propose answers to several other questions which may, at first glance, appear to be unrelated: What is the best way to characterize the frequency distribution of scientific productivity? How do scientific disciplines differ in productivity? Why are the mean and standard deviation of productivity distributions so closely related? What is the appropriate functional form relating productivity to other independent variables?

Disciplinary Differences in the Scale of Productivity

I shall begin by examining differences in productivity across disciplines. Table 1 presents the means and standard deviations of publications and citations for a sample of scientists teaching in graduate departments in the US in 1965.[9] If we take the standard deviation as our measure of inequality, Table 1 clearly shows that (a) citations are much more unequally distributed than publications, and (b) for both citations and publications, chemistry shows the greatest inequality, followed in order by biology, physics, and mathematics. Notice, however, that an increase in the standard deviation is always accompanied by an increase in the mean. In fact, for publications the two statistics have nearly the same numerical values. This suggests that the disciplinary differences in Table 1 merely reflect differences in the *scale* of productivity, rather than differences in equality.

TABLE 1
Means and Standard Deviations of Publications and Citations for Scientists Teaching in US Graduate Departments

	Mathematics	Physics	Biology	Chemistry
Articles Published, 1961-65				
Mean	5.11	7.43	10.94	11.56
Standard Deviation	6.42	8.42	11.01	13.50
N	361	499	507	555
Citations to Lifework, 1966				
Mean	7.02	15.85	21.82	28.67
Standard Deviation	20.51	39.63	46.03	75.28
N	453	554	495	628

Let me explain this idea by way of an example. Suppose that the mathematicians in the sample divided each of their papers into two shorter papers and published each of them separately, thus doubling their 'productivity' as measured by article counts. This would have the effect of doubling the mean to 11.22, almost that of the chemists. It would also double the standard deviation to 12.83 which is, again, almost the same as the chemists' standard deviation. Most people would agree, I think, that this change is purely artifactual and does not represent a real increase in the inequality of productivity among the mathematicians. We have merely changed the scale of productivity since what were formerly half papers are now counted as full papers. Incidentally, this example is not so far-fetched as it might seem. There is evidence that papers in mathematics journals are about twice as long, in number of words, as papers in chemistry journals. Moreover, biology and physics fall between chemistry and mathematics in number of words per paper.[10]

What is special about this example is that differences in the scale of productivity are artifactual. But more generally, we can say that two distributions differ only in scale whenever one distribution can be obtained from the other by raising (or lowering) each person's score by a fixed percentage. Such a change will always have the effect of raising (or lowering) both the mean and the standard devia-

tion by the same percentage. In the case of artifactual changes of scale, it is clearly desirable to have a measure of inequality that does not depend on the scale of productivity. But even when scale changes represent real increases (or decreases) in productivity there are several reasons for preferring inequality measures that are scale invariant.[11] Perhaps the most intuitive reason is that increasing everyone's score by the same percentage does not change each person's share of the total number of publications. Thus if six percent of the scientists publish 50 percent of the papers and we increase everyone's productivity by 100 percent, it will still be true that six percent of the scientists publish 50 percent of the papers.

Most of the familiar measures of dispersion like the range, the standard deviation, and the average deviation, respond directly to scale changes. As we have seen, doubling everyone's score doubles the standard deviation. There are, however, several measures of inequality which are invariant to changes of scale. These include the Gini index, the coefficient of variation, the standard deviation of the logarithms and an information-theoretic measure proposed by Theil.[12] All scale invariant measures are functions of individuals' percentage shares of the total number of publications rather than their absolute number of publications. Hence, it can be said that they respond to relative differences among individuals rather than to absolute differences. Elsewhere I have discussed the relative merits of these measures in considerable detail.[13] While each has advantages and disadvantages that depend on the application, the coefficient of variation appears to have particular merit when applied to counts of articles and citations.

One standard of comparison is sensitivity. Theil's measure of inequality and the standard deviation of the logarithms are most sensitive to differences at the lower end of the distribution, and their sensitivity diminishes as one goes up the scale. For J-shaped distributions, the Gini index behaves in the same way.[14] To make this notion of diminishing sensitivity more concrete, suppose that we take five publications away from one scientist and give them to another scientist. This will produce a greater change in the standard deviation of the logarithms if both scientists are low producers than if both are high producers. Diminishing sensitivity is often thought to be desirable in measuring inequality of income because it corresponds to the principle of diminishing marginal utility. But in the absence of any theory or evidence about diminishing marginal utility of publications or citations, it seems preferable to choose a

measure of inequality that is equally sensitive to differences at all levels of the distribution.[15] The coefficient of variation has this property. Moreover, it is by far the easiest to compute since it is just the standard deviation divided by the mean. Finally, as I shall show in the next section, the coefficient of variation is easily modified to correct for purely random variation in the distribution of publications and citations.

Note that it is usually irrelevant whether one uses the coefficient of variation or its square — the variance divided by the squared mean — since one is an increasing function of the other. For the remainder of the paper it will be more convenient to use the squared coefficient of variation, which I will often refer to as simply 'the coefficient of variation'. This is *not* equivalent, however, to Gaston's 'coefficient of variability', which is the variance divided by the mean.[16] Since Gaston's measure is not scale invariant, the validity of his results on increasing inequality is questionable.

Computing squared coefficients of variation for the data in Table 1, we get the following results: for publications, mathematics 1.58, chemistry 1.32, physics 1.28, biology 1.01; for citations, mathematics 8.54, chemistry 6.90, physics 6.25, biology 4.40. These results corroborate two conclusions reached using the standard deviation as the measure of inequality: citation inequality is much higher than publication inequality, and the rank order for publications is the same as the rank order for citations. But the rank order itself has changed drastically. Mathematics, which previously had the lowest inequality, now has the highest, while the lowest position is now occupied by biology.[17]

We see, then, that the mean and the standard deviation appear to give redundant information about the distribution of productivity because both are primarily determined by the scale of productivity. Taking their ratio, however, cancels out the effect of scale and gives us rather different information about the level of inequality. Notice that, unlike the Gini index which is bounded by zero and one, the coefficient of variation has no upper bound. Some people feel more comfortable with the Gini index for this reason, but it is hardly a valid basis for preferring one measure over the other. In fact, since the intuitive feel of the Gini index comes largely from transferring notions of what constitutes a high correlation or a high probability, the upper bound of one can be more misleading than informative.

The Modified Coefficient of Variation

When the variable of interest is the number of events that occur in some interval of time, all the scale invariant measures suffer from a peculiar problem — they tend to increase as the interval of time becomes shorter. This phenomenon is illustrated in the upper panel of Table 2 which gives squared coefficients of variation for articles published by 239 chemists in the years 1965 through 1969.[18] I first computed the coefficients for each of the four years, then summed the counts into two-year counts and recomputed the measures. Finally, I computed the coefficient of variation for the combined four-year counts. Since the coefficient of variation is scale invariant, one would expect that the level of inequality would be about the same whether computed for four-year totals or for single years. We find, however, that inequality for the two-year counts is lower than either of the one-year coefficients for the same years, and that the four-year inequality is less than either of the two-year coefficients. In fact, the two-year inequality averages 20 percent

TABLE 2
Inequality of One-Year, Two-Year, and Four-Year Counts of Articles, 239 Chemists

Squared coefficient of variation

1965	2.44	2.13	2.03
1966	3.32		
1967	3.99	2.78	
1968	3.23		

Modified squared coefficient of variation

1965	1.47	1.59	1.75
1966	2.07		
1967	2.79	2.20	
1968	2.10		

above the four-year level, and the one-year inequality averages 60 percent more than the four-year coefficient. This could be a serious problem if one wished to compare groups whose publications had been counted over intervals of different length.

Why does this occur and what can be done about it? What appears to be happening is that over short intervals of time there is a great deal of purely random variation in the number of articles a scientist publishes or the number of citations received. Even if a scientist submitted one paper exactly every six months, the contingencies of the publication process would virtually ensure that in some years he would have no publications, while in other years he might have three or four. Such random variation tends to increase the standard deviation without increasing the mean and, hence, is reflected in a larger coefficient of variation. Over longer intervals, however, the random variation tends to cancel out and has less impact on the coefficient of variation and other measures of inequality.

To deal with this problem, it is necessary to postulate a statistical model for the occurrence of the random variation. One plausible model is well-known to statisticians as a mixture of Poisson processes. I have described the model in detail elsewhere,[19] but I will summarize it here. Let us suppose that each scientist has a number R which is his latent or unobserved rate of publication (or rate of being cited). For instance, R might be 1.62 papers per year. The actual, observed number of publications in an interval of length t is allowed to vary randomly with a mean value of Rt. The simplest assumption that can be made about this random variation is that it follows a Poisson distribution with the parameter equal to Rt. Since every scientist has a different value of R, however, the probability distribution of publications for all scientists will not be a Poisson distribution but a mixture of Poisson distributions. (This notion will be elaborated in the next section.)

Since the observed number of publications is assumed to contain purely random variation, it would be more desirable to know the coefficient of variation for R, the latent rate of publication. Even though R cannot be directly observed, its coefficient of variation can be expressed in terms of the mean and variance of the observed number of publications. This modified, squared coefficient of variation, which I will denote by C, is given by

$$C = \frac{\sigma^2 - \mu}{\mu^2}$$

where σ^2 is the variance and μ is the mean number of publications or citations. The only difference between C and the unmodified coefficient of variation is that the mean is substracted from the variance in the numerator. Although C is not scale invariant for the observed number of publications, it *is* scale invariant for R, the latent rate of publication.

The lower panel of Table 2 shows the effect this has on the comparisons made earlier. All the modified coefficients are, of course, lower than the unmodified coefficients. But the one-year coefficients are reduced much more than the two-year coefficients, which in turn are reduced more than the four-year coefficients. As a consequence, the relationship between interval length and inequality is much less pronounced. There is still some tendency for the coefficients to be larger in the shorter intervals, but at least the inequality for the aggregated counts now falls within the range of the disaggregated counts.

The Distribution of Productivity

A further advantage of the modified coefficient of variation C is its close relationship to a theoretical distribution which fits the observed distribution of productivity. In 1926 Lotka proposed his now-famous inverse square law of productivity.[20] If x is a given number of publications, the inverse square law says that the proportion of scientists publishing x papers is given by

$$p(x) = \frac{6}{(\pi x)^2}$$

Although Lotka's distribution provides a moderately good fit for samples drawn from abstracting services, it has some drawbacks. First, it is mathematically intractable since its theoretical mean is infinite and its variance is undefined. Second, it often provides a poor fit to data produced by first drawing a sample of scientists independently of the abstracting service, and then counting how many papers each has published over some fixed interval of time.[21] For this sort of data, much better fits are usually obtained with the negative binomial distribution, a discrete probability distribution which is well known in statistics, has desirable mathematical properties, and follows from several different theoretical models, two of which are relevant here. A summary of its properties is given by

TABLE 3
Observed and Expected Frequencies for the Distribution of Article Counts Assuming a Negative Binomial Distribution

	Chemists, 1967-68		Biochemists, Career Years 1-3	
Number of Articles	*Observed*	*Expected*	*Observed*	*Expected*
0	129	126.4	59	59.6
1	30	38.2	65	61.6
2	23	21.5	46	50.1
3	13	14.0	35	37.1
4	11	9.7	23	26.1
5	10	7.0	28	17.8
6	5	5.1	11	11.9
7	9		9	7.8
8	1	6.8	3	5.1
9	1		3	
10	1	5.3	2	5.3
11	1		1	
12	1		0	3.5
13	2	5.0	0	
14	1		0	
15	0		0	
16	0		0	
17	0		0	
18	1		0	
23	0		1	
$\bar{x}$	1.716		2.545	
s^2	6.670		8.196	
χ^2		5.96		8.64
d.f.		7		8
p		0.54		0.36

Bishop and his colleagues, but for a thorough exposition see Johnson and Kotz.[22]

Table 3 shows the fit of the negative binomial distribution to two-year counts of articles for the 239 chemists described in the previous section. The fit is quite good, either by a visual inspection or by a chi-square goodness-of-fit test. Similar fits were obtained for various aggregations of the four-year productivity data described earlier. Table 3 also gives similar results for 286 biochemists whose articles were counted in the first three postdoctoral years.[23] Neither of these two distributions was well described by Lotka's inverse square law.

Of course, just because the negative binomial fits these distributions well does not mean that other theoretical distributions would not do an equally good job. What makes the negative binomial so attractive is that it can be derived as a consequence of two theoretical models which can be plausibly applied to scientific productivity. The first is simply an elaboration of the mixture of Poisson processes discussed in the last section. Recall that each scientist has a latent rate of publication R, but his actual number of papers is drawn from a Poisson distribution with a parameter equal to Rt. We assumed earlier that different scientists had different values of R but we said nothing about the shape of the distribution of R. If we now assume that R has a distribution which is a member of the gamma family, it follows that the observed distribution of publications should be negative binomial.[24] The gamma family is a large class of unimodal distributions which have an origin at zero. Its members include the chi-square distribution and the exponential distribution, and it is often used as an approximation when the precise distribution is unknown.

The second model was introduced to sociologists by Coleman who referred to it as a 'contagious Poisson process'.[25] It is more commonly known as a 'linear birth process', or 'linear birth with immigration'. The model assumes that all scientists start out with the same value of R, the latent rate of publication, but each time a scientist publishes a paper his rate of publication is incremented by some constant amount. I have described this model in detail elsewhere.[26] Allison and Krauze have used the model as a starting point for building models of cumulative advantage.[27] The essential idea is that a variety of social reward or feedback processes imply that each additional publication increases the likelihood of future publications.

The nice thing about being able to characterize a frequency distribution by a theoretical probability distribution is that one can summarize the entire distribution by a few parameters. For example, knowing the mean and variance of a normal distribution tells you all there is to know about that distribution. In the case of the negative binomial, the probability distribution is given by

$$p(x) = \left[\frac{k+x-1}{x}\right]\left[\frac{k}{k+m}\right]^{k}\left[\frac{m}{k+m}\right]^{x}$$

where p(x) is the probability that a scientist will publish x papers. There are two fixed parameters in this formula — m, which is just the theoretical mean or expectation, and k.[28] This brings us to the main point of this section. The modified squared coefficient of variation is equal to 1/k. Hence, when one has estimated C, one has a fundamental parameter of the distribution of productivity. This is a further justification for C as a measure of inequality of scientific productivity.

Efficient Estimation of C

In Table 2, I estimated C in a rather obvious way. I took the sample variance, subtracted the sample mean, and divided by the squared sample mean. This is a good rough and ready method, but there are better methods. In statistical terminology, this simple method — known as 'the method of moments' — is inefficient; that is, it tends to fluctuate more around the true value than do other, efficient methods. This is particularly troublesome if one is fitting distributions as in Table 3, since the validity of the chi-square goodness-of-fit test depends on the assumption that the parameters of the distribution are efficiently estimated. The problem also becomes more acute when the true value of C becomes large, as it does with counts of citations. This fact will be illustrated shortly.

An alternative method is maximum likelihood, which produces estimators which are optimally efficient in large samples. This method requires an interative procedure which I have described elsewhere.[29] Assuming that the distribution of productivity is, in fact, negative binomial, one first gets the estimate of the parameter k and then takes its reciprocal to get the maximum likelihood

estimate of C. A FORTRAN programme for the interative estimation of k is available on request.

To demonstrate the difference that the estimation procedure can make, I first computed the maximum likelihood estimate of C for citations received in 1966 by the 239 chemists, obtaining a value of 4.089. When I used the simpler method based only on the sample mean and the sample variance I got an estimate of 7.317, almost 80 percent larger. Table 4 shows what happens when the reciprocals of these values (that is, the k parameter) are used to generate predicted values of the negative binomial distribution. The predictions generated from the maximum likelihood estimate are reasonably close to the observed figures, and based on these estimates we cannot reject the negative binomial distribution by the chi-square test. On the other hand, the predicted values based on the simpler method-of-moments are so far off that a chi-square test is hardly necessary. In general, the efficiency of the method-of-moments estimator decreases as either the population mean increases or the population value of C increases. When both C and μ are less than one, the moment estimator will give quite good estimates as compared with maximum likelihood. Otherwise, the maximum likelihood estimator has a substantial advantage.[30] My own experience is that the method-of-moments estimator and the maximum likelihood estimator are usually quite close for publication distributions, but that they differ markedly for citation distributions.

The Form of the Production Function

As a final note, let us return to the problem of accounting for the differences in the means and standard deviations of publications and citations across four disciplines. I suggested that these differences were essentially differences in scale — that the effect of being in chemistry rather than, say, physics is to raise one's productivity by a certain fixed percentage. Note that this differs from the usual analysis of variance model which says that the effect of being in a particular category is to *add* a constant to each person's score. In the scale model, we *multiply* each person's productivity by an amount that is constant within each category.

TABLE 4
Expected and Observed Frequencies for Counts of Citations Received by 239 Chemists in 1966

Number of Citations	Observed	Expected Based on Maximum Likelihood	Expected Based on Method of Moments
0	102	107.7	141.7
1	36	25.3	18.9
2	16	15.2	10.5
3	15	10.9	7.3
4	9	8.5	5.6
5	6	6.9	4.6
6	6	5.8	3.8
7	6	5.0	3.3
8-9	11	8.2	5.4
10-12	4	9.2	4.3
13-17	7	10.5	3.5
18-24	6	9.0	3.0
25-33	5	6.7	2.5
$\geqslant 34$	10	10.0	24.5
$\bar{x}$	6.117		
s^2	279.92		
χ^2		13.26	67.0
d.f.		11	11
p		0.28	< 0.001
C		4.089	7.317

In symbols, this multiplicative model can be expressed as

$$Y_{ij} = \theta \alpha_j \omega_{ij}$$

where Y_{ij} is the productivity of person i in discipline j, α_j is the 'effect' of being in discipline j, ω_{ij} is a random disturbance and θ is the geometric mean of the expected productivity scores.

Taking the logarithm of both sides transforms this multiplicative model into one that is both linear and additive:

$$\log Y_{ij} = \log \theta + \log \alpha_j + \log \omega_{ij}$$

This, in turn, can be rewritten as

$$\log Y_{ij} = \beta + \varphi_j + \epsilon_{ij}$$

It is plausible that other variables also affect productivity, not by adding a given amount to scientists' productivity but by increasing or decreasing productivity by a certain percentage. This model can be generalized to include such variables by writing:

$$\log Y_{ij} = \beta_o + \varphi_j + \beta_1 x_{i_1} + \ldots\ldots + \beta_k x_{ik} + \epsilon_{ij}$$

A reasonable way to estimate this model is to take the logarithm of the productivity score and use this logged value as the dependent variable in a multiple regression. This approach has the additional virtue of constraining the predicted values to be non-negative, as we know they must be. However, a problem arises in that there will usually be some people who publish nothing or receive no citations, and the log of zero is undefined. The usual recourse is to add a small constant to everyone's score. But what constant to use? Statistical reasoning leads to the conclusion that log $(Y + \frac{1}{2})$ produces the least biased results.[31]

The logarithmic transformation of publication and citation counts has been used by others, but the usual rationale (if one is given) is that taking the logarithm reduces the skewness of the distribution and makes it more closely approximate a normal distribution.[32] This is a dubious justification since (contrary to popular opinion) the assumptions of multiple regression do not require that the observed variables be normally distributed. Here, however, we have a theoretical rationale for the logarithmic func-

tional form — namely, that independent variables tend to produce multiplicative increments or decrements in productivity rather than additive increments or decrements.

Conclusion

Inequality measures are usually chosen on purely technical grounds, but the criteria I have applied here involve theoretical issues as well. For example, the choice between scale invariant and scale dependent measures will depend on whether one is more concerned with relative differences (in which case a scale invariant measure should be chosen) or with absolute differences among individuals. The functional form of the relationship between productivity and other variables is also relevant to this decision. In choosing among the various scale invariant measures, the key question is whether one thinks that differences among low producers are more important than differences among high producers, or whether differences at all levels should be equally weighted. I have also shown that the theoretical probability distributions of productivity, and models for the generation of that distribution, are relevant for the choice of an inequality measure. On all these criteria, I have suggested that a modified coefficient of variation is the best measure of the inequality of scientific productivity.

• NOTES

I am indebted to Barbara F. Reskin and Warren O. Hagstrom for the use of their data reported here. The biochemistry data in Table 3 were collected in collaboration with J. Scott Long and Robert McGinnis under contract N01-0D-4-2175 with the NIH Office of Program Planning and Evaluation.

1. D. J. de S. Price, *Little Science, Big Science* (New York: Columbia University Press, 1963), Ch. 2.

2. J. P. Britton, 'The Productivity of Scientists: A Prelude to Manpower Studies' (unpublished paper, Department of History of Science and Medicine, Yale University, November 1964).

3. P. D. Allison and J. A. Stewart, 'Productivity Differences among Scientists: Evidence for Accumulative Advantage', *American Sociological Review*, Vol. 39 (1974), 596-606.

4. R. K. Merton, 'The Matthew Effect in Science', *Science*, Vol. 159 (5 January 1968), 56-63; S. Cole and J. R. Cole, *Social Stratification in Science* (Chicago: The University of Chicago Press, 1973); Allison and Stewart, op. cit. note 3.

5. S.Cole, J. R. Cole and L. Dietrich, 'Measuring the Cognitive State of Scientific Disciplines', in Y. Elkana et al. (eds), *Toward a Metric of Science* (New York: Wiley, 1978), 209-51.

6. M. A. Faia, 'Productivity among Scientists: A Replication and Elaboration', *American Sociological Review*, Vol. 40 (1975), 825-29.

7. J. Gaston, *The Reward System in British and American Science* (New York: Wiley, 1978), 148-51.

8. Allison and Stewart, op. cit. note 3; Cole et al., op. cit. note 5.

9. For a detailed description of this sample, see W. O. Hagstrom, 'Competition and Teamwork in Science', Final Report to the National Science Foundation for Research Grant GS-657, 1967, or Hagstrom, 'Competition in Science', *American Sociological Review*, Vol. 39 (1974), 1-18.

10. National Science Foundation, Office of Science Information Service, *Characteristics of Scientific Journals, 1949-59* (Washington, DC: National Science Foundation, 1964).

11. For a more extensive discussion see P. D. Allison, 'Measures of Inequality', *American Sociological Review*, Vol. 43 (1978), 865-80.

12. H. Theil, *Economics and Information Theory* (Chicago: Rand McNally, 1967), 92.

13. Allison, op. cit. note 11.

14. Ibid.

15. While it might be argued that additional publications have decreasing marginal utility to the individual scientist in terms of social rewards that follow, it is not so easy to maintain that there is diminishing marginal utility to the scientific community. See P. D. Allison, *Processes of Stratification in Science* (PhD dissertation, University of Wisconsin, Madison, 1976), 73-74; in press, same title (New York: Arno Press, 1980).

16. Gaston, op. cit. note 7, 148.

17. The same rank ordering is obtained with the Gini index. See Allison and Stewart, op. cit. note 3.

18. For a description of the sample, see B. F. Reskin, 'Scientific Productivity and the Reward Structure of Science', *American Sociological Review*, Vol. 42 (1977), 491-504.

19. P. D. Allison, 'Reliability of Variables Measured as the Number of Events in an Interval of Time', in K. F. Schuessler (ed.), *Sociological Methodology 1978* (San Francisco, Calif.: Jossey-Bass, 1977), 238-53.

20. A. J. Lotka, 'The Frequency Distribution of Scientific Productivity', *Journal of the Washington Academy of Sciences*, Vol. 16 (1926), 317.

21. Allison, op. cit. note 15, 94.

22. Y. M. M. Bishop, S. Fienberg and P. Holland, *Discrete Multivariate Analysis* (Cambridge, Mass.: MIT Press, 1975), 454-56; N. L. Johnson and S. Kotz, *Discrete Distributions* (Boston, Mass.: Houghton Mifflin, 1964).

23. For a description of the sample, see J. S. Long, 'Productivity and Position in the Scientific Career', *American Sociological Review*, Vol. 43 (1978), 889-908.

24. S. Spilerman, 'The Causes of Racial Disturbances: A Comparison of Alternative Explanations', *American Sociological Review*, Vol. 35 (1970), 627-49.

25. J. S. Coleman, *Introduction to Mathematical Sociology* (New York: The Free Press, 1964), 299.

26. P. D. Allison, 'Estimation and Testing for a Markov Model of Reinforcement', *Sociological Methods and Research*, Vol. 9 (1980), forthcoming.

27. P. D. Allison and T. K. Krauze, 'The Effect of Cumulative Advantage on Inequality in Science', paper presented to the American Sociological Association annual meeting, Chicago, September 1977.

28. When k is an integer, the first term on the right hand side is the binomial coefficient $(k+x-1)!/(k-1)!x!$. When k is not an integer the generalized binomial coefficient is $\Gamma(k+x)/\Gamma(x+1)\Gamma(k)$ where $\Gamma(.)$ is the gamma function.

29. Allison, op. cit. note 26.

30. H. S. Sichel, 'The Estimation of the Parameters of a Negative Binomial Distribution with Special Reference to Psychological Data', *Psychometrika*, Vol. 16 (1951), 107-51.

31. D. R. Cox and P. A. W. Lewis, *The Statistical Analysis of Series of Events* (London: Methuen, 1966), 43-44. An alternative estimation procedure is that of maximum likelihood under the assumption that each person's score has a Poisson distribution. This is easily accomplished with the programme GLIM, distributed by the Numerical Algorithms Group, 7 Banbury Road, Oxford OX2 6NN, UK.

32. W. O. Hagstrom and L. L. Hargens, 'Mobility Theory in the Sociology of Science', paper presented at the 63rd Annual Meeting of the American Sociological Association, Boston, Mass., 27 August 1968; Hagstrom, 'Inputs, Outputs, and the Prestige of University Science Departments', *Sociology of Education*, Vol. 44 (1971), 375-97; F. Clemente, 'Early Career Determinants of Research Productivity', *American Journal of Sociology*, Vol. 79 (1973), 409-19.

Paul Allison is Assistant Professor of Sociology at Cornell University. His major interests are stratification and productivity in science and statistical methods for the analysis of categorical data. Currently, with J. Scott Long, he is studying the relationship between characteristics of university departments and the research productivity of their members. His recent publications include *Processes of Stratification in Science* (New York: Arno Press, 1980) and 'Analyzing Collapsed Contingency Tables Without Actually Collapsing, *American Sociological Review*, Vol. 45 (February 1980). *Author's address*: SASS, 323 Uris Hall, Cornell University, Ithaca, New York 14853, USA.