ISSN: 2281-1346

Department of Economics and Management

DEM Working Paper Series

H Index: A Statistical Proposal

Paola Cerchiello
(Università di Pavia)

Paolo Giudici
(Università di Pavia)

# 39 (04-13)

Via San Felice, 5
I-27100 Pavia
http://epmq.unipv.eu/site/home.html

April 2013

H INDEX: A STATISTICAL PROPOSAL

Paola Cerchiello and Paolo Giudici

Department of Economics and Management, University of Pavia

Corresponding Author: Paola Cerchiello, via S. Felice 5, 27100 Pavia, paola.cerchiello@unipv.it

Abstract

The measurement of the quality of academic research is a rather controversial issue. Recently Hirsch has proposed a measure that has the advantage of summarizing in a single summary statistics all the information that is contained in the citation counts of each scientist. From that seminal paper, a huge amount of research has been lavished, focusing on one hand on the development of correction factors to the h index and on the other hand, on the pros and cons of such measure proposing several possible alternatives. Although the h index has received a great deal of interest since its very beginning, only few papers have analyzed its statistical properties and implications, typically from an asymptotic viewpoint. In the present work we propose an exact statistical approach to derive the distribution of the h index. To achieve this objective we work directly on the two basic components of the h index: the number of produced papers and the related citation counts vector, by introducing convolution models. Our proposal is applied to a database of homogeneous scientists made up of 131 full professors of statistics employed in Italian universities. The results show that while "sufficient" authors are reasonably well detected by a crude bibliometric approach, outstanding ones are underestimated, motivating the development of a statistical based h index. Our proposal offers such development and in particular exact confidence intervals to compare authors as well as quality control thresholds that can be used as target values.

1 Foreword

The measurement of the quality of academic research is a rather controversial issue. Recently Hirsch (2005) has proposed a measure that has the advantage of summarizing in a single summary statistics all the information that is contained in the citation counts of each author. From that seminal paper, a huge amount of research has been lavished, focusing on one hand on the development of correction factors to the h index (Iglesias and Pecharroman 2007, Burrell 2007, Glanzel 2006) and on the other hand, on the pros and cons of such measure proposing several possible alternatives (Todeschini, 2010, and others therein).

Concerning the first stream of research, Glanzel in 2006 analyzed the basic mathematical properties of the h index thanks to the adoption of the Paretian distribution for the citation count, stressing the strength of such index when the available set of papers is small (that is the case for young researchers mainly). Iglesias and Pecharroman in 2007 proposed to use a simple multiplicative correction to the h index able to take into account the differences among researchers coming from different science citation index (SCI) fields and thus allowing a fair and sustainable comparison. Indeed these authors offer a table with such normalizing factors according to specific distributional assumptions of the citation counts (power law or stretched exponential model). Burrell in 2007 made a step ahead since he proposed to employ a stochastic model for an author's production/citation patterns. In that framework it is possible to consider different situations according to the level of production and citation or the length of a researcher's career.

Very prosperous is the literature focusing on possible alternative to the h index, in particular we cite Todeschini that in 2010 reviewed such proposed indexes (g index, hg index, e index, A index, R index, R_m index, r_w index, h_w index, $h^{(2)}$ index, W_u index etc..) and proposed a new one, the j index, based on the h index formula plus a correction term needed to take into account the excess of the publications in the h index core and the distribution of the citations. Todeschini (2010) also makes a final comparison assessment among all such indexes, employing a classical statistical multivariate method based on principal component analysis.

Although the h index has received a great deal of interest since its very beginning (see e.g. Ball 2005), only two papers have analyzed its statistical properties and implications: Beirlant and Einmahl (2010) and Pratelli et al. (2012). Beirlant and Einmahl demonstrated the asymptotic normality of the empirical h index for the Pareto-type and Weibull-type distribution families, allowing the construction of asymptotic confidence intervals of each author and evaluating the statistical significance of the difference between two authors with the same academic profile (in terms of career length and SCI field.) Very recently Pratelli et al. (2012) investigated, in a full statistical perspective, the distributional properties of the h index and the large sample expressions of its relative mean and variance, in a discrete distributional context.

In the present work, expanding the seminal contribution of Glanzel (2006) we propose an exact, rather than asymptotic, statistical approach. To achieve this objective we work directly on the two basic components of the h index: the number of produced papers and the related citation counts vector. Such quantities will be modelled by means of a compound stochastic distribution, that exploits, rather than eliminate, the variability present in both the production and the impact dimensions of a scientist's work.

The paper is organized as follows: in section 2 we present our proposal; in section 3 we apply the new approach to a database of scientists homogeneous by career age and scientific culture, finally, section 4 contains some concluding remarks.

2 Proposal

The measurement of the research achievements of scientists has received a great deal of interest, since the paper of Hirsch (2005) that has proposed a "transparent, unbiased and very hard to rig measure" (Ball, 2005): the h index. The information needed to calculate the h index of a scientist is contained in the vector of the citation counts of the N_p papers published by a scientists along her/his career.

The Hirsch definition is that "a scientist has index h if h of his or her N_p papers have at least h citations each and the other (N_p-h) papers have $\leq h$ citations each".

Following the seminal work of Hirsch, many papers have dwelled on this issue, especially in the bibliometric community. Surprisingly, few papers have focused on the statistical aspects behind the h index, apart from Glanzel (2006) that hinted at the relevance of a "statistical background" for the h index. Recently Beirlant and Einmahl (2010) and Pratelli et al. (2012) have proposed an asymptotic distribution for the h index that can be used for inferential purposes and not only for descriptive summaries, as in the typical bibliometric contributions. Our contribution follows such recent papers, with the aim of providing a statistical framework for the h index that, in addition, holds also for small sample sizes and respects the discrete nature of the bibliometric data at hand.

Let $X_1, \ldots, X_n$ be random variables representing the number of citations of the N_p articles (henceforth for simplicity n) of a given scientist. We assume that $X_1, \ldots, X_n$ are independent with a common citation distribution function F. Beirlant and Einmahl (2010) and Pratelli et al. (2012), among other contributions, assume that F is continuous, at least asymptotically, even if citation counts have support on the integer set.

According to this assumption, the h index can be defined in a formal statistical way as in Glanzel (2006) and Beirlant and Einmahl (2010):

$$h : 1 - F(h) = \frac{h}{n}$$

A different statistical definition can be found in Pratelli et al. 2012:

$$h = sup\{x \geq 0 : nS(x) \geq x\}$$

where

$$S(x) = P(X > x)$$

is the survival function and

$$\bar{S}(x) = P(X \geq x)$$

is its left-hand limit.

From our point of view, the definition should be as much as possible coherent with the nature of the data and, therefore, in the present paper we assume that F is discrete and, in order to define the h index, we employ order statistics.

Given a set of n papers of a scientist to which a citations count vector $\underline{X}$ is associated, we consider the ordered sample of citations $\{X_{(i)}\}$, that is $X_{(1)} \geq X_{(2)} \geq \ldots \geq X_{(n)}$, from which obviously $X_{(1)}$ ($X_{(n)}$) denotes the most (the least) cited paper. Consequently the h index can be defined as follows:

$$h = max\{t : X_{(t)} \geq t\}$$

The main latent assumption behind all the above mentioned definitions is the adequacy of the h index as a summary measure. For example,Hirsch claimed that the h index is better than other measures such as the total number of citations because the latter ”may be inflated by a small number of big hits” (Hirsch, 2005). However, from a proper statistical viewpoint, the h index is not a sufficient statistics, as will be shown in the following.

In order to prove this, we need to derive the exact distribution of the h index itself. Order statistics can be profitably employed for this purpose, as in the procedure outlined in Cerchiello and Giudici (2012), as in the following.

The exact distribution of the h index is:

$$p(h_i) = [F(X_i) - F(X_{i-1})]^{(n+1-h_i)}$$

Using the above distribution the h index can be shown not to be sufficient. For example, assume a scientist has produced two papers, with citations counts X_1, X_2. Assume that X_1 and X_2 are an i.i.d. sample from a Poisson distribution of parameter λ. We want to check whether the h index $H = f(X_1, X_2)$ is a sufficient statistics for (X_1, X_2). Consider the set of possible sample events as follows:

Now, if we calculate the conditional probability function $Prob(X_1 = 1, X_2 > 1|H = 1)$ and, for simplicity and without loss of generality, we assume that $X_2 = 3$, we obtain:

Table 1:

	(0,0)	(0, $X_2 > 0$)	($X_1 > 0$, 0)	(1, 1)	(1, $X_2 > 1$)	($X_1 > 0$,1)	(2, 2)	(2, $X_2 > 2$)	($X_1 > 2$, 2)	($X_1 > 2$, $X_2 > 2$)
H	0	1	1	1	1	1	2	2	2	2

$$Prob(X_1 = 1, X_2 = 3|H = 1) = \frac{\frac{e^{-2\lambda}\lambda^{(1+3)}}{1!3!}}{\frac{(e^{-\lambda}\lambda^{1})^{(2+1-1)}}{1!}} = \frac{\lambda^2}{6}$$

From the above formula is evident that the conditional probability function depends on the parameter λ, thus we conclude that the h index statistics H is not a sufficient summary statistics.

On a constructive side, a sufficient statistics for the citation vector $\underline{X}$ may be the total number of citations and its bijective functionals. The total number of citations has not been considered a valid summary by Hirsch because of his high sensitivity to outlying observations. Although this may be a questionable remark, it can be naturally taken into account in an appropriate statistical framework as in the model that we are going to propose.

Consider a setting in which the majority of observations have a small probability of occurrence and few ones have a large one. This is a typical situation in loss data modeling (see e.g. Cruz, 2002). In this context the number of occurrences of a specific event, n, is a discrete random variable and the loss impact of each occurrence is another random variable (typically continuous) conditional on the former. The two distributions can then be compounded deriving the distribution of the total impact loss. Note that such loss data model takes obviously into account both large probability/small impact and small probability/high impact events.

The logic behind loss data models can be extended to the evaluation of research impact of a scientist or of a community of scientists, and this is our proposal. This requires interpreting the number of occurrences as the number of papers produced along the career of a scientist and the vector of impacts as the vector of citation counts of the papers of the same scientist.

References to statistical models for loss data modeling can be found in the so-called Loss Distribution Approach (LDA) (see for example Cruz, 2001 and Dalla Valle and Giudici, 2008) where the losses are categorized in terms of 'frequency' and 'severity' (or impact). The frequency is the random number of loss events occurred during a specific time frame, while the severity is the mean impact of all such events in terms of monetary loss.

In our context the frequency is the (random) number of published papers along the career of a scientist and the impact is the (random) mean number of citations received in the same time frame by all such papers. Let $X_i = (X_{i1}, X_{i2}, \ldots, X_{in_i})$ be a random vector containing the citations of the n_i papers published by the i-th scientist. Note that, not only X_i but also n_i is a random quantity that can be denoted with the term 'frequency'. Consequently, the total impact of a scientist i can be defined as the sum of a random number n_i of random citations:

$$C_i = X_{i1} + X_{i2} + \ldots + X_{in_i}$$

Note that the above formula can be equivalently expressed as follows:

$$C_i = n_i \times m_i$$

where $m_i = \frac{\sum_{j=1}^{n_i} C_{ij}}{n_i}$ is the mean impact of a scientist.

Our aim is to derive the distribution of the sufficient statistics C_i and of functionals of interest from it that can be interpreted as statistical based research quality measures, such as the h index, $H_i = f(C_i)$. In order to reach this objective one additional assumption has to be introduced.

We assume that, for each scientist $i = 1, \ldots, I$ in a homogeneous community, conditionally on the production of each scientist (with number of papers equal to n_i), the citations of the papers X_{ij}, for $j = 1, \ldots, n_i$ are independent and identically distributed random variables, with common distribution $k(x_i)$:

$$k(x_{i1}) = k(x_{i2}) = \ldots = k(x_{in_i}) = k(x_i)$$

On the basis of the previous assumption we can derive the distribution of the total number of citations C_i of each scientist, through the convolution of the frequency distribution with the paper citations distribution that are therefore the building components of our proposed approach.

For each scientist i, the distribution function of C_i, that is $F_i(x) = P(C_i \leq x))$, can thus be found by means of a convolution between the distributions of n_i and m_i as follows:

$$F_i(x) = \sum_{n_i=1}^{\infty} p(n_i) k^{n_i *}(x_i)$$

where is the $k^{n_i *}$ indicates the n_i-fold convolution operator of the distribution $k(.)$ with itself (see e.g. Buhlmann 1970 and Frachot et a. 2001):

$$k^{1*}(x_i) = k(x_i)$$
$$k^{n*}(x_i) = k^{(n-1)*}(x_i) * k(x_i)$$

and, for each scientist, $p(n_i)$ is the distribution of the number of produced papers and $k(x_i)$ is the distribution of the paper citations.

In practice, the distribution functions $p(n_i)$ and $k(x_i)$ depend on unknown parameters, say λ_i and θ_i. A reasonable modeling assumption is that n_i, the number of published papers of a scientist in a specific community, follows a distribution $p(n_i|\lambda_i)$ with λ_i a parameter that summarizes the productivity of each scientist and that, conditionally on n_i, the paper citations x_i follows a distribution $k(x_i|\theta_i, n_i)$ with θ_i a parameter that is function of the mean impact that may vary across scientists. While it is reasonable to take $\lambda_i = \lambda$, especially for a population with common characteristics in terms of seniority and scientific publication behavior, θ_i is unlikely to be constant. For example, θ_i can vary according to the number of published papers (as in Iglesias and Pecharroman, 2005, Burrell, 2007); this implies letting $\theta_i = \theta * n_i$. A different way to model over dispersion is to let θ_i follow a $Gamma(\alpha, \beta)$

distribution. This leads to a negative binomial distribution whose estimate however requires the knowledge of the total number of papers produced worldwide that can potentially cite the papers of the scientists under analysis.

We remark that a more trivial assumption could be to model directly the total number of citations C_i, for example as a Poisson distribution: this simplistic assumption evidently discards the fact that the total number of citations is function of individual paper citations each of which may have a different distribution. This effect can be incorporated in our convolution model.

To complete the proposed model we need to specify two parametric distributions, one for the production and one for the citation patterns.

For example, a starting assumption may be to take:

$$p(n_i|\lambda_i) \sim Poisson(\lambda_i)$$
$$k(x_i|\theta_i, n_i) \sim Poisson(\theta_i)$$

where λ_i and θ_i are unknown and strictly positive parameters to be estimated, representing, respectively, the mean number of published papers and the mean number of citations of each scientist (the mean impact).

Under the above assumption, the maximum likelihood estimates of the two parameters can be easily seen to be:

$$\hat{\theta} = \frac{S}{N}$$
$$\hat{\lambda} = \frac{N}{I}$$

where $N = \sum_{i=1}^{I} n_i$, $S = \sum_{i=1}^{I} \sum_{j=1}^{n_i} C_{ij}$.

Once parameters are estimated the distribution functions of C_i and $H_i = f(C_i)$ can be obtained and quality measures can be derived. From the distribution of H_i one can calculate appropriate statistical summaries that can be used for inferential purposes on science achievements. For example, the top 5% percentile of the distribution represents a high quality threshold for a given community of scientists; the percentile associated with a specific H_i may instead be taken to describe the quality ranking of that scientist in the community; finally, for each scientist the point estimate corresponding to the observed H_i can be supplemented with a confidence interval.

However the above summaries and, more generally, functional of interest from $F_i(x)$ may not be obtained analytically. In this rather frequent case one can resort to Monte Carlo simulations to approximate numerically $F_i(x)$. Our approach can thus provide a natural inferential framework for the estimation of the h index which is not, differently from Pratelli et al. (2012), based on large sample assumptions.

The starting Poisson-Poisson assumption can be modified so to obtain a better fit to the data. For the distribution of the number of papers, we have observed that, in communities characterized by a high level of heterogeneity in the production process, a discrete uniform distribution may be more appropriate. Conversely, as far as citations are concerned, what observed by Hirsch (the h index may be inflated by very few papers with a large number of citations) can be embedded into a discrete extreme value distribution, such as the Zipf-Mandelbrot distribution (see e.g. Mandelbrot 1962, Evert et al. 2004, Izack, 2006), that parallels continuous EVT distributions such as the Pareto (as in Glanzel, 2006).

Specifically, we assume that the ordered citation counts of each scientist $X_{i(j)}$ are associated with ranks $r_{i(j)}$ that follow a Zipf-Mandelbrot distribution (hereafter ZM):

$$f(r_{i(j)}) = \left(\frac{T}{r_{i(j)} + \beta}\right)^{\alpha} for \quad r_{i(j)} = 1, \dots,$$

where for a given scientist i, α is parameter that describes the decay rate of the ranks distribution, β is a smoothness parameter and finally T is a normalizing constant. According to the support of the rank positions $r_{i(j)}$ we can have two versions of the Zipf-Mandelbrot distribution:

- Zipf-Mandelbrot with infinite support (ZM): in this case $r_{i(j)}$ has no upper bound;

- Zipf-Mandelbrot with finite support (fZM): in this case $r_{i(j)}$ is finite, albeit large, with support $r_{i(j)} = 1, \ldots, S$, thus we have an extra parameter that is S.

In the next section we show the results obtained by employing both ZM and fZM distributions, as well as a uniform distribution for the number of papers.

3 Application

We consider a database of homogeneous scientists made up of 131 full professors of statistics employed in Italian universities. This is indeed a small subset of the worldwide population of statisticians. These scientists forms a cohort of people that has grown their careers under similar conditions: both in terms of academic rules (they belong to the same country) and in terms of research 'modus operandi' (they belong to the same scientific community). To our knowledge this is the first time in bibliometric studies that a community of scientists rather that single top scientists have been considered in the analysis.

Such database has been collected by a public organization named 'VIA-Academy' (www.via-academy.org) that aims at improving the quality of Italian scientists by providing open feedbacks on their research quality on a bibliometric basis. The database, that has been sent to us, contains the Google-scholar based h indexes of all Italian statisticians updated until 1st June 2011. We have cleaned and refined the data, and added for each scientists, her/his citation counts vector. The refinement has involved a long activity of disambiguation (from homonimies and wrong affiliations) that was carried out by employing the well known 'Publish or Perish' (Harzing, 2007).

Finally, as said before, we have selected only full professors to guarantee a homogeneous cohort of scientists especially with regards to career time. For a discussion of time effects on bibliometric indexes see e.g. Hirsch 2005, Beirlant et al. 2010.

Figure 1 describes our data in terms of observed total number of citations for the considered scientists.

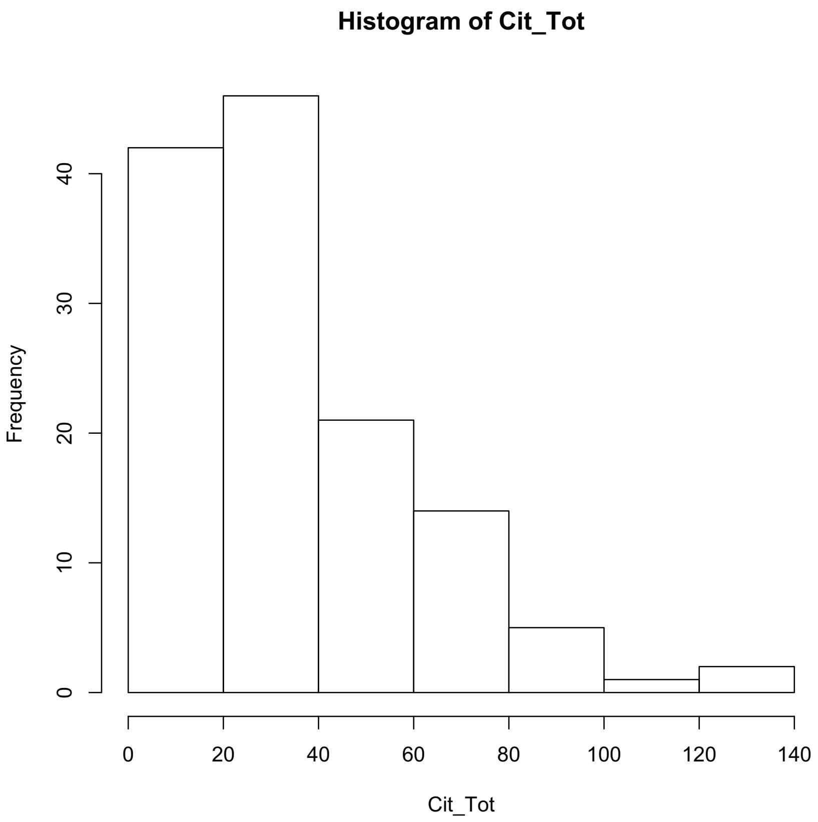
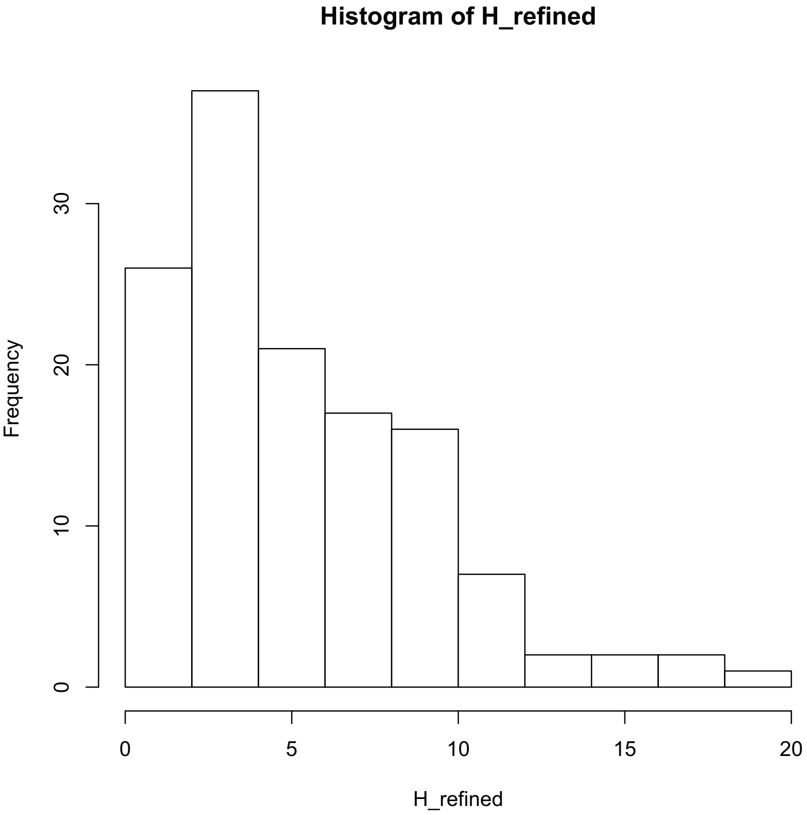


Figure 1 about here

From Figure 1 note that the distribution of citations is, as expected, right skewed. Indeed, from the above distribution the main summary statistics assume the following values: mean=206.2, median=107, maximum=2438 and minimum=0.

In Figure 2 we report for the same data the distribution of the h index.

Figure 2 about here



From Figure 2 note that the distribution of the h index is, as expected, similar to the distribution of the citations in Figure 1, but appears more concentrated: mean=5.68, median=5, maximum=19 and minimum=0. Indeed the skewness and kurtosis of the h index are 0.98 and 0.90, instead for the distributions in Figure 1 are 3.82 and 20.53.

So far we have considered a bibliometric analysis of our data. We now move to a proper statistical framework. As already remarked, statistical inference can shed more light on the issue of finding an appropriate tool to measure scientific quality.

The distribution functions $p(n_i)$ and $k(x_i)$ will be estimated from our data that can be thought as of a sample of scientists assumed with common citation distribution $F_i(x)$.

The observed sample correlation between the number of published papers and the total citations impact is equal to 0.62, and therefore we explore the case $\theta_i = \theta * n_i$, in addition to the simpler assumption $\theta_i = \theta$.

Specifically for the publication rate generating mechanism, we have considered two alternative distributions: a Poisson distribution with mean parameter equal to $\lambda = 35.66$ (the observed mean number of published papers for the full professors at hand) and a discrete Uniform distribution with parameter N equal to the maximum number of papers produced ($N = 128$).

Concerning the citation rate generating mechanism, we have considered four alternative distributions, conditional on the number of produced papers, n_i: first a Poisson distribution with mean parameter equal to 4.67 (the observed mean number of citations among the scientists at hand) then a Poisson distribution with mean parameter equal to 4.67 multiplied by n_i/N, with n_i the observed number of papers for the i-th scientist and N defined as before. We have then considered a Zipf-Mandelbrot distribution with decay parameter rate equal to 0.51, an estimate obtained with the log-log regression method suggested by Gabaix (2009), times the same correction factor as for the Poisson distribution. We remark that the observed correlation between log citations and log ranks is indeed equal to 0.8 and this justifies the application of the log-log estimation method (see e.g. Gabaix, 2009). Finally we consider a finite Zipf-Mandelbrot distribution, since it is reasonable to assume that the random variable describing the rank of the papers $ri(j)$ has a finite support. Following the method implemented in R software, package zipfR (Evert, 2007, Evert and Baroni, 2004), we obtained parameter estimates equal to $\alpha = 0.065$, $\beta = 5.6e - 05$ and $S = 51.75$.

We have thus considered eight convolutions: Poisson-Poisson ($\theta_i = constant$), Poisson-Poisson ($\theta_i = variable$), Poisson-ZM, Poisson-fZM, Uniform-Poisson, Uniform-ZM, Uniform-fZM. In addition we have considered a simple Poisson distribution for the citations.

We have compared all the above eight distributions in terms of a chi-squared goodness of fit test that is based on the difference between the observed frequency distribution of the h index and the expected one, under each of the seven alternative modelizations. For the comparison we have employed throughout ten intervals of equal size, resulting in 9 degrees of freedom. The application of the comparison to our data shows that the Uniform-Poisson convolution leads to the best fit, with $\chi^2 = 0.4166$, followed by the Uniform-fZM convolution with $\chi^2 = 2.3024$ both clearly significant. The remaining convolutions are not significant.

The difference between the Uniform-Poisson and the the Uniform-fZM convolutions is not so evident and can depend on the data at hand and on the characteristics of the set of considered professors.

On the basis of the best fit distribution (the Uniform-Poisson), a quality threshold of 90% is equal to $h = 8$. This can be taken by the considered community as a target value to be overcome by those who look for high quality results. On the other hand, a threshold of 50%, often taken as a minimum requirement in career competitions (see e.g. www.anvur.it) is equal to $h = 5$. These results should be compared with the bibliometric values, obtained directly from the observed distribution of the h index, which are equal, respectively, to: $h = 11$ and $h = 5$. This means that, while "sufficient" authors are reasonably well detected by a crude bibliometric approach, outstanding ones are underestimated, because of a too heavy weight of "big hits" which is exactly what Hirsch' index proposed to remediate. Our statistical based threshold overcomes this problem, setting a higher threshold.

We now consider the application of what proposed to the comparison of individual scientists, considering in particular three top scientists in the community. We considered either the Uniform-Poisson and the Uniform-fZM convolutions to evaluate the most performing approaches that can be different from the previous context since the citation vector is now referred to a specific author. For the parameters of the Uniform random variable, we propose to calculate the deciles of the number of papers distribution on the whole dataset and we consider the minimum and maximum values of each decile intervals. For the citations vector mechanism, we consider both a Poisson and a fZM distribution. We have considered as running example, the top performing authors in the community of all Italian full professors of statistics: Mr. X (rank 1), with an observed h index of 19 and 128 papers, Mr. Y (rank 2), with an observed h index of 17 and 123 papers and Mr. W (rank 4) with an observed h index of 12 and 78 papers (all updated at May 2011). For each of them we have estimated the Uniform-Poisson and the Uniform-fZM convolutions estimating the parameters on the relative citations vectors. It turns out that the parameters of the uniform distribution are the same (a=78, b=128) for each author; the mean number of papers is equal to $\lambda = 8.51$

(Mr. X), $\lambda = 11.98$ (Mr. Y) and $\lambda = 2.30$ (Mr. W) and the decay parameter is equal to $\alpha = 0.64$ (Mr. X), $\alpha = 0.54$ (Mr. Y) and $\alpha = 0.57$ (Mr. W).

In order to quantify the real difference among Mr. X , Mr. Y and Mr. W we can now calculate the confidence intervals of their h index with level of confidence equal to 90%. Table 1 shows the results:

Table 2: Confidence intervals for the h index for Mr. X, Mr. Y, Mr. W under the Uniform-Poisson (U-P) and the Uniform-fZM (U-fZM) distributions.

Scientist	U-P	U-fZM
Mr X (observed h=19)	[11;13]	[19; 29]
Mr Y (observed h=17)	[13;15]	[12; 20]
Mr W (observed h=12)	[6;8]	[11; 17]

From Table 1 the reader can infer that the Uniform-Poisson convolution is unable to capture the empirical h index, always underestimating the real value. On the other hand the Uniform-fZM convolution contains the real value and moreover clearly shows that the two top ranking scientists are not significantly different between each other since their corresponding confidence intervals overlap. Finally Mr. W, even if belongs to the same percentile (the top 10%), is significantly different from the first top scientist but not from the second.

4 Conclusions

In this paper we address the topic of evaluating the quality achievements of scientists taking statistical variability into proper account. The well known Hirsch index (the h index) is convincing from a purely bibliometric perspective but not from a stochastic viewpoint. We overcome this problem by embedding the citation counts, of which the h index is a function, in an appropriate probability framework that takes inspiration from loss data modeling.

The resulting 'statistical h index' can thus boost the descriptive power of the measure proposed by Hirsch, not limiting it to summary purposes but allowing inferential evaluations, as in the recent paper by Pratelli et al. 2012. The added value of our proposal is not to rely on the large sample distribution of the h index but to fully respect the discrete nature of the data by deriving the exact distribution of the h index and proposing a discrete convolution model to draw exact inferential conclusions.

From an applied perspective, we foresee at least two main advantages in the adoption of our statistical h index:

1. comparison among scientists and among different communities of scientists can be robustified by using appropriate confidence intervals and levels;

2. scientific quality can be monitored by devicing appropriate statistical quality control thresholds.

Indeed our approach can be applied not only to compare scientists but also to the comparison of scientific institutions and research departments, to the comparison of scientific communities and to the comparison of research time periods without loss of generality.

Moreover, our proposal can be profitably applied to contexts different from the evaluations of scientists, characterized by two types of information that can be summarized by a random variable representing a count frequency and a random variable representing the corresponding impact.

BIBLIOGRAPHY

Beirlant, J., and J. H. J. Einmahl, 2010, Asymptotics for the Hirsch Index: Scandinavian Journal of Statistics, v. 37, p. 355-364.

Ball, P. 2005, Index aims for fair ranking of scientists, Nature, 436:900.

Burrell, Q. L., 2007, Hirsch's h-index: A stochastic model: Journal of Informetrics, v. 1, p. 16-25.

Cerchiello, P., and P. Giudici, 2012, On the distribution of functionals of discrete ordinal variables: Statistics Probability Letters, v. 82, p. 2044-2049.

Cruz M. G. 2002, "Modeling, measuring and hedging operational risk." Wiley.

Dalla Valle, L., and P. Giudici, 2008, A Bayesian approach to estimate the marginal loss distributions in operational risk management: Computational Statistics Data Analysis, v. 52, p. 3107-3127.

Evert, S. 2004, A simple LNRE model for random character sequences. In Proceedings of the 7mes Journes Internationales dAnalyse Statistique des Donnes Textuelles (JADT 2004), pages 411422, Louvain-la-Neuve, Belgium

Evert, S. and Baroni, M. 2007, zipfR: Word frequency distributions in R. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Posters and Demonstrations Session, Prague, Czech Republic.

Gabaix, X., 2009, Power Laws in Economics and Finance: Annual Review of Economics, v. 1, p. 255-293.

Glanzel, W., 2006, On the h-index - A mathematical approach to a new measure of publication activity and citation impact: Scientometrics, v. 67, p. 315-321.

Harzing, A.W. 2007 Publish or Perish, available from http://www.harzing.com/pop.htm

Hirsch, J. E., 2005, An index to quantify an individual's scientific research output: Proceedings of the National Academy of Sciences of the United States of America, v. 102, p. 16569-16572.

Iglesias, J. E., and C. Pecharroman, 2007, Scaling the h-index for different scientific ISI fields: Scientometrics, v. 73, p. 303-320.

Izsak, F., 2006, Maximum likelihood estimation for constrained parameters of multinomial distributions - Application to Zipf-Mandelbrot models: Computational Statistics Data Analysis, v. 51, p. 1575-1583.

Mandelbrot, B. 1962, On the theory of word frequencies and on related Markovian models of discourse. In R. Jakobson (ed.), Structure of Language and its Mathematical Aspects, pages 190219. American Mathematical Society, Providence, RI.

Pratelli, L., A. Baccini, L. Barabesi, and M. Marcheselli, 2012, Statistical Analysis of the Hirsch Index: Scandinavian Journal of Statistics, v. 39, p. 681-694.

Todeschini, R., 2011, The j-index: a new bibliometric index and multivariate comparisons between other common indices: Scientometrics, v. 87, p. 621-639.