Understanding Citation Indices

Some love them, some hate them, but citation indices are heartily gobbled up by administrators in tenure and promotion decisions. It has also been argued that funding should be tied to citation history (Nicholson and Ioaniddis, Nature, 2012). Adler, Ewing, and Taylor (Stat. Sc., 2009), and Hall (Stat. Sc., 2009) are among the best expositions for statisticians. We now have bountiful of citation indices. The h-index due to *Hirsch*, the g-index of *Egghe*, and the quite recent $i10$ index introduced by *Google* are among the most chic. My h-index is k if my most cited k papers have each been cited k or more times, but the next most cited paper didn't get $k+1$ citations. So, for example, if Mr. Smith's citation numbers are $2000, 200, 100, 30, 20, 20, 7, 7, 2, 0$, then his h-index is seven. My g-index is k if my most cited k papers have been cited k or more times on the average; so, Mr. Smith's g-index is 10. The $i10$ index is simply the number of articles with 10 or more citations, the idea being that ten citations means it got looked at.

The pros and cons of the indices have been greatly discussed. Citation indices across fields simply cannot be compared. To be in the top 1% of most cited physicists, you must have 2073 citations; the same number in interdisciplinary sciences is 147 citations. I do not want to spend my page talking about these again. My intention is to try to rigorously understand scales for these indices so I can make personal judgements of a reported value, average, low, or high? Can we understand this at a level higher than a purely empirical level? Perhaps we can. I must however refer to Pratelli et al. (SJS, 2012) and references therein for supremely powerful different insights.

My one page forces me to look here at just one index, and I choose h. If someone has n publications, and the order statistics of the citations, arising from a CDF F with density f, are $X_{(1)}, \ldots, X_{(n)}$, then h equals the largest k such that $X_{(n-k+1)} \geq k$. It is useful to write this in terms of the quantile function of the data; the h-index is $n \sup\{0 < t < 1 : F_n^{-1}(1-t) - nt \geq 0\}$. We center and normalize to get the *quantile process* $q_n(t) = \sqrt{n} f(F^{-1}(t))[F_n^{-1}(t) - F^{-1}(t)]$, and then, $h = n \sup\{0 < t < 1 : q_n(1-t) \geq n^{3/2} t f(F^{-1}(1-t)) - \sqrt{n} f(F^{-1}(1-t))F^{-1}(1-t)\}$. With very carefully imposed conditions on f, the quantile process is approximated (even strongly) by a BB (Brownian Bridge) on $[0, 1]$. Since time reversal is ok for a BB, we have that h is roughly equal in law to

$n(1-\tau)$ where τ is the hitting time of a curved boundary $a(t)$ by a BB. This is one hard nut to crack exactly analytically.

Fortunately, some of the best minds have looked at this; let me just mention Borovkov, Daniels, Durbin. Of special utility today is Durbin (JAP, 1985) (although see Durbin (JAP, 1992) with David Williams's rejoinder), where a sequence of approximations to the density of the hitting time is spelled out. The first approximation is about the only one I can write on paper, but it already gives ample insight to h. If we look at researchers of all ages then it is very hard to fathom the citations as being iid from one F. I guess that at the assistant professor level, F could be something like a uniform on $[0, m]$ for smallish m. At the associate level, it'd be already skewed, perhaps an exponential; for senior people, it will be extremely skewed. For ten of us at the full professor level who work on theory, I find the largest citations to be $528, 207, 643, 750, 708, 498, 601, 69, 38, 31$; folded Cauchy comes to mind. For the uniform case, on following Durbin's theorems, we arrive at a beautiful answer: Durbin's first approximation to the density for $\tau/(1-\tau)$ is an inverse Gaussian density (of independent fame in random walk theory, e.g., Feller), with parameters $\mu = n/m$ and $\lambda = n^3/m^2$, so that the mean of $\tau/(1-\tau)$ is n/m and its variance is $1/m$. Notice that the variance is small. At a casual level, this says that our h-index would be about $1/(1/m + 1/n)$ on the average. We are beginning to see a theoretical benchmark, not empirical alone. For promotion to associate, we may ask for 12 publications with some of them having 15 citations, so average performane would require an h of $1/(1/12 + 1/15) = 7$ for theoretical researchers. In general, the density of τ itself would be *bimodal*; this comes out of more calculations from Durbin's formula. The bimodality can be useful; if you are past the higher mode, you certainly deserve promotion!

Of course, I simulated to test my math. The inverse normal calculation predicts that for a uniform citation pattern on $[0, 30]$ with $n = 35$, h should be confined between 11 and 16. I simulated six times(!); and h varied between 9 and 14. I can sleep well with that agreement. The g-index will involve the running integral of a BB, which too can be handled, and the $i10$ will only require hitting time by a Brownian motion, which is classic (e.g., Lerche (Springer, 1986)). The case of the nonlinear boundary

will require another column, if we still have any passion left. I wonder if instead of a single number, we can look at a five number summary, say $(X_{(n)}, X_{(n-1)}, \text{median citations per article}, \text{percentage of papers with} < 5 \text{ citations}, \text{and } n)$. The h-index of course changes as time passes; in a sense that can be made rigorous, it acts like a nonhomogeneous Poisson process. We do not have room to explain. Now, if someone can confirm that my h is more than 7, I think I will then rightfully demand tenure.