IMU-ICIAM-IMS 報告 Citation Statistics

国際数学連合（IMU）応用数理国際評議会(ICIAM) および数理統計学会（IMS）により標記の報告書（注）が公表されました。これは、研究評価における被引用データの利用および誤用に関する報告書です。

日本数学会と日本学術会議数理科学委員会では、上記の団体より許諾を頂き、この報告書を共同で翻訳致しました。その全文を「数学通信」誌上および数学会のホームページ（注）で公表致します。

実際の翻訳作業は，小田忠雄氏（東北大学名誉教授）に行って頂きました。また、翻訳に際して、赤平昌文氏（筑波大学）および竹村彰通氏（東京大学）より有益な助言を頂きました。これらの方々に深く感謝致します。

尚、この日本語訳は自由に引用または利用して頂いて結構ですが、その場合は出典を明示して頂きますようお願い致します。

（注）次の URL にアクセスすると、報告書の原文（英語）と日本語訳へのリンクが表示されるページへ行きます。 http://mathsoc.jp/IMU/

（担当理事 岩崎克則記）

引用統計

定量的研究評価に関する
国際数学連合(IMU, International Mathematical Union),
応用数理国際評議会
(ICIAM, International Council of Industrial and Applied Mathematics),
数理統計学会
(IMS, Institute of Mathematical Statistics)
合同委員会報告

ロバート・アドラー(Robert Adler),
(委員長)ジョン・ユーイング(John Ewing),
ピータ・テーラー(Peter Taylor)

(2008年6月11日,2008年6月12日訂正版)

要旨

本報告書は，研究評価に際する引用データの利用および誤用に関する報告書である．研究評価に際しては「簡単かつ客観的な」方法を使用すべきであるとの見方がますます有力になっている．「簡単かつ客観的な」方法とは，**文献計量**(ビブリオメトリックス)，即ち引用データおよびそれに基づく統計であるとの解釈が一般的である．複雑な判断を単純な数字に置き換えるが故に，引用データは本来的に正確であり，ピア・レビュー(同分野の専門研究者による査読)に際する主観性の弊害を排除すると信じられている．しかしながらそれは根拠薄弱である．

- 統計の使い方が不適切であれば，統計の信頼性は，データの信頼性より高くはあり得ない．誤用・誤解された統計により誤った結論に至ることもあり得るのである．現代文献計量学の多くは，引用統計の解釈と正しさに関するこれまでの経験と直観に依存しているようである．
- 数字は一見「客観的」ではあるが，客観性は錯覚にすぎないこともあり得るのである．引用がピア・レビューよりもかなり主観的なことさえあるのである．引用データの主観性があまり明白ではないが故に，引用データの使用者はその限界に気付きにくい．
- 引用データのみに頼ってしまうと、最良の場合でも，研究と言うものの不完全で表面的な理解に留まってしまい，他の判断によって補強されてのみ正しい理解となるのである．**数字は，正しい判断よりも本来的に優るというものではない．**

引用データを使用した研究評価とは，結局のところ，引用データを使った学術誌・論文・研究者・プログラム・学術分野のランク付けを行うことを意味する．しかし，これらのランク付けの際に使用される統計ツールは，しばしば誤解・誤用されている．

- 学術誌に関しては，ランク付けのためにインパクト・ファクターがもっとも頻繁に使用される．これは当該学術誌に掲載された論文の被引用状況の分布の単純な平均である．この平均は，その分布に関するごくわずかな情報しか捉えておらず，大雑把な統計に過ぎない．それに加え、被引用状況で学術誌を評価する際には混乱を招く要因がいろいろ存在するので，インパクト・ファクターを使用して学術誌を比較する場合には注意を要する．インパクト・ファクターのみで学術誌の評価を行うことは、ある人の健康を体重のみで判断するようなものである．
- 論文を比較するに際して，各論文の被引用回数ではなく，論文の掲載された学術誌のインパクト・ファクターを代わりに使用する場合がしばしばある．インパクト・ファクターが大きいほど、被引用回数が多いであろうと考えるのであろうが，

多くの場合**間違い**である！ 蔓延しているこの統計の誤用に対しては，見付け次第異議申し立てをしなければならない.

- 研究者個人の被引用状況を網羅的に比較するのは困難なため，研究者の被引用状況と言う非常に複雑なものを単一数値として把握できるような簡単な統計手法がいくつか試みられてきた．その中で最も有名な h 指数の人気はますます高まっているようであるが，一見しただけでも，h 指数とその類似指数は，複雑な被引用状況を理解しようとするナイーブな試みであることが判る．ある研究者の被引用状況に関する僅かな情報しか捉えておらず，研究評価の際に枢要な情報を見失っている.

インパクト･ファクターや h 指数のような指数の有効性は､まだよく判っておらず，研究されてもいない．時には「経験」に基づいて，これらの指数と研究の質とが関連するとされる．「簡単に入手可能」という理由から、これらの指数に依拠することを正当化している．これまでに行われた数少ない研究も，被引用データから有用な情報をいかにして最大限に引き出すかというよりも，質に関する他の指標との相関を示すことに焦点を絞った狭いものに過ぎない.

我々は､研究評価の手段としての引用統計を否定しているわけではない．被引用データと統計は価値ある情報をある程度は提供できると考える．評価は実用的なものでなくてはならず，従って簡便に得られる引用統計が評価に際して一定の役割を演じるであろうことは我々も認識している．しかしながら，被引用データは，研究の質に関してほんの限られた不完全な視点しか提供できないし，被引用データから得られた統計量は､時として十分理解されず，誤用もされている．研究とは極めて重要なものであって，粗っぽい手段一つで簡単に価値を計れるようなものではない.

評価に関わっておられる読者の方々に本報告の解説および詳説をお読み頂き，引用統計の限界とより有効な利用法を御理解頂くよう切に願っている．研究遂行に高い規範を設定するのなら，その質の評価にも高い規範を設定するべきであろう.

定量的研究評価に関する
国際数学連合(IMU),
応用数理国際評議会(ICIAM),
数理統計学会(IMS)
合同委員会

ロバート・アドラー(Robert Adler, イスラエル工科大学テクニオン),
(委員長)ジョン・ユーイング(John Ewing, アメリカ数学会),
ピータ・テ―ラー(Peter Taylor, メルボルン大学)

委員会付託事項より

学界における透明性と説明責任を一層求める結果、統計データのアルゴリズム的評価によって公平な決定が達成できるのだと個人や機関が信じる「数字の文化」が生み出された．意思決定者は，質を計測するとの究極の目標が不可能であるため，計測可能な数字に置き換えた．この動向に対しては，「数字を扱う」プロである数学者や統計学者から一言あってしかるべきである．

序

科学研究は重要である．研究は，近代世界の発展の基礎であり，我々人類が直面する環境問題から人口爆発問題に至るまでの，解決不可能かと思われるさまざまな課題の解決策を見出せるかもしれないとの希望を抱かせてくれる．そのため，世界中の政府や機関が科学研究に膨大な財政的支援を提供しているが，注ぎ込んだ資金が賢明に投資されているかどうかを知りたがっている．資金を提供している研究の質を評価し，将来の投資決定に役立てようと考えるからである．

これは何も新しいことではない．長年にわたって研究評価は行われてきた．では、いったい何が**新しい**のかと言えば，良い研究評価とは「簡単かつ客観的」でなければならず，しかも，科学者自身の判断も含む様々な方法ではなく，被引用データから得られる統計量に主として依拠することにより実現が可能と言う考え方である．最近発表された某報告書は，冒頭でこの考え方を明確に述べている．

> 英国研究評価(Research Assessment Exercise)において、大学における研究の質的評価に現在使われている方法は，次期サイクル終了時の2008年以後は別の方法に変更すべきであると政府は考えている．ピア・レビューではなく計量が新システムの焦点となり，学術誌の文献やその被引用回数を使用する文献計量がこの新システムにおいて質を計る中心的指標となると期待している．[エビデンス社報告 2007, Evidence Report 2007, p.3]

簡単な客観性の推進派は，研究は極めて重要であって，主観的判断に頼るだけでは済まされず，被引用に基づく計量によってランク付けの透明性が高まり，他の評価法に内在する曖昧性を排除出来ると信じているのである．また、注意深く選択された計量は偏見の影響を受けず独立性が高いと信じている．その上，そのような計量によって，研究に関わるすべて，即ち，学術誌・論文・研究者・プログラム，さらには学術分野全体までも，主観的ピア・レビューに依らずに単純かつ効率的に比較することができると考えているのである．

しかしながら，文献計量の正確さ，独立性，効率性に対する信仰は見当はずれである．

- 第一に，文献計量の正確さは幻想に過ぎない．統計を誤用することで嘘をつくことができることは常識である．引用統計の誤用が蔓延している．そのような誤用(例えば、インパクト・ファクターの誤用)に対する度重なる警告にも関わらず，政府、機関、研究者自身さえもが，引用統計の誤用に基づく不当なあるいは間違った

結論を導き出し続けている.

- 第二に，被引用に基づく計量のみに依拠することは，結局のところ，ある種の判断を別のものに置き換えているに過ぎない．主観的ピア・レビューを，引用の主観的解釈に置き換えているに過ぎない．被引用に基づく計量のみに依存することの推進派は，研究が引用される毎の意味が、当該研究の「インパクト」という同一のものであると暗黙の内に仮定している．この仮定は検証されていず，間違っている可能性が高い.
- 第三に、我々が生きている世界を理解するのに統計は有用ではあるが、統計が提供できるのは部分的な理解に過ぎない．数値的計量が他の理解方法よりも優れているとの神秘的思想を提唱することが，現代社会において一種の流行になっている．研究を全面的に理解することの**代替品**として引用統計の使用を推奨する人々は，暗黙の内にこのような神秘的思想を持っている．我々は統計を**正しく**使うだけでなく，**賢明に**使う必要がある.

研究を評価することを問題にしているのではない．研究評価には主として「簡単で客観的な」引用計量に依るべきであるとの要求を問題にしているのである．この要求は，出版物や研究者やプログラムを順位づけるための計算しやすい数値を使用すべきとしているものと解釈されるのがしばしばである．研究には，短期的目標と長期的目標があるのが普通であり，従って，研究の価値は複数の基準に基づいて判定されるべきである．二つのものが必ずしも比較できないという意味で，全順序のつけられないものが現実世界にも抽象世界にも沢山あることは，数学者には常識である．比較には，複雑な分析がしばしば必要であり，二つの内のどちらが「より優れている」かが決められないことも起こりうる.「どちらが優れているか？」との質問に対する正解が，「場合によりけり！」と言うこともあるのである.

研究には複数の目標があるのが普通であり，従って，研究の価値は複数の基準に基づいて判定されるべきである.

研究の質的評価に複数の方法を使用して欲しいとの要請は以前にも行われた（例えば，[Martin 1996]や[Carey - Cowling - Taylor 2007]）．公表された研究は，被引用状況以外にも様々な方法で評価できる．招待講演や編集委員委嘱や受賞などの名声でも質を計ることができる．分野や国によっては，研究助成金を受けていることも評価に役立つ．同分野の研究者による評価であるピア・レビューも評価に際して重要である．（誤用される欠点があるからとの理由で引用統計を切り捨てるべきではないのと同様に，偏見が入り得るからとの理由でピア・レビューを切り捨てるべきではない.）上記は，多様な評価方法のほんの一例に過ぎない．良い評価に至るには様々な道があり，それらの内のどれが重要かは，学問分野によって異なる．それにも拘わらず，被

引用状況に基づく「客観的な」統計が度々優先される．処理の簡単さや簡単な数値(出来得れば単一数値)であることの誘惑が，常識や賢明な判断を曇らせてしまうようである．

研究評価に際する統計の誤用をなんとかしようとの意図の下に，本報告は数理科学者によって纏められた．統計の誤用では数学分野が被害を被っていることも，本報告をまとめるに至った理由の一つである．学術誌・論文・著者の被引用回数が少ないという数学分野特有の文化の所為で，数学は引用統計誤用の被害を受けやすい．しかし、**すべての**研究者と一般大衆が，研究評価に際して賢明な科学的手法を使用するよう努力すべきであると信ずる．

一部の研究者は，過去に誤用されたとの僻みから，引用統計を切り捨ててしまうことを望んでいるが，貴重なツールを切り捨ててしまうことになるので賛成できない．もし正しく使用され，注意深く解釈され，さらに一連の処理の一部なのであれば，被引用に基づく統計は研究評価に役立ち**得る**．被引用状況は，学術誌・論文・研究者に関する情報を提供する．我々はその情報を隠蔽したいのではなく，むしろ光を当てたいのである．

それが報告書の目的である．冒頭の3節では，学術誌･論文･研究者の評価に際して，被引用データがどのように使用(または誤用)され得るかを扱う．その次の節では，引用の意味するところが如何に多様であるか，従って，被引用状況に基づく統計にはどのような限界があるかについて述べる．最後の節では，統計を賢明に使って頂きたいこと，また，評価に際して，作業が多少複雑になることを厭わず，引用統計を他の判断でぜひ調整して頂きたいことも提案する．

アルバート･アインシュタインは，「何事も出来得る限り単純化しなければならないが，必要以上に単純化してはならない」と述べた．(注1) 世界で最も卓越した科学者の一人である彼のこの忠告は，研究評価にまさにぴったりである．

学術誌のランク付け：インパクト・ファクター (注2)

インパクト･ファクターは，一定期間内での論文当たり平均被引用回数により学術誌の価値を計る方法として，1960年代に創られた［Garfield 2005］．この平均値はJournal Citation Reports 誌の発行元であるトムソン・サイエンティフィック社（Thomson Scientific，従来は，科学情報研究所 Institute for Scientific Informationと呼ばれていた．）［訳注：2008年現在Thomson Reuters トムソン・ロイター社]によって収集

されたデータから計算される．トムソン・サイエンティフィック社は9,000以上の学術誌[訳注: 原文では，以下でindexed journalsと称しているものであるが．本稿では以下で「データ収録誌」と訳する]から，各論文の書誌とその論文が引用している全文献の書誌とで構成する引用情報を抽出し，毎年データベースに追加している［THOMSON: SELECTION］．この情報を使えば，特定の論文が，データ収録誌掲載の論文に何回引用されたかを計ることができる．（トムソン・サイエンティフィック社の使用するデータ収録誌の数は，数学における主要二次情報誌 Mathematical ReviewsとZentralblattの使用するデータ収録誌の数の半分以下であることに留意したい.(注3)）

特定の学術誌の特定の年に関するインパクト･ファクターとは，当該学術誌に当該年より前の2年間に掲載された論文が,（トムソン・サイエンティフィック社のデータ収録誌に)当該年に掲載された論文で引用された平均回数として計算する．ある学術誌の2007年のインパクト・ファクターが1.5であるとは，当該誌に2005年と2006年に掲載された論文が，2007年にデータ収録誌に掲載された論文で平均して1.5回引用されたことを意味する．

トムソン･サイエンティフィック社自身も，データ収録誌の選定基準の一つとしてインパクト・ファクターを使用している［THOMSON: SELECTION］．他方では，もっと一般に学術誌を比較する際にインパクト･ファクターを使用することを，次のように奨励している．

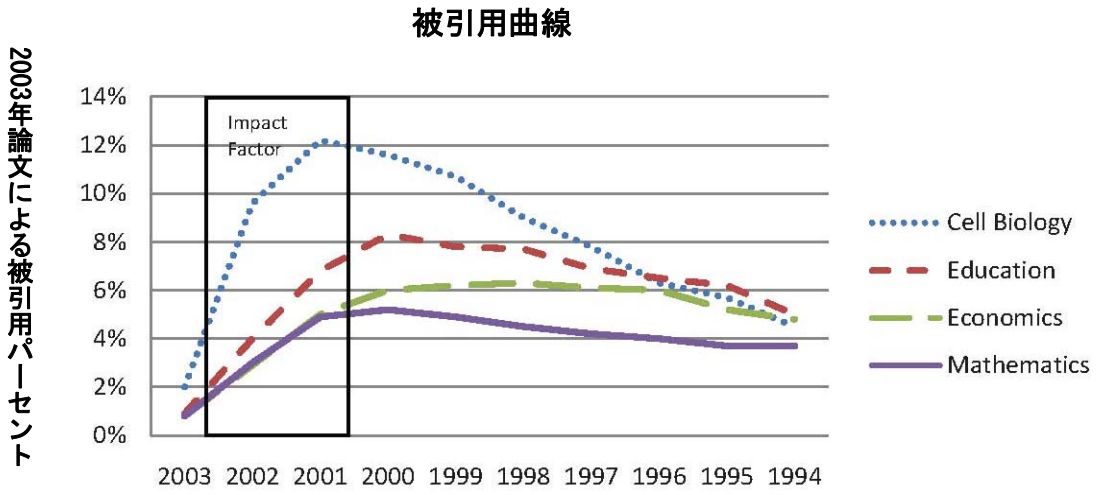
> 「インパクト･ファクターは，図書館における購読済み学術誌や購読検討中の学術誌に関する情報を，図書館職員に提供する学術誌管理ツールであるが，学術誌購読に際しては，費用や利用状況に関するデータも併用して合理的な決定を下すべきである.」［THOMSON: IMPACT FACTOR］

被引用データのみで学術誌の学術的価値を評価するべきではないと，これまで多くの筆者が指摘しており，本報告書の著者一同もその意見には強く同意する．このような一般的な観察に加え，インパクト・ファクターは他の理由によっても批判されてきた.（[Seglen 1997]，[Amin - Mabe 2000]，[Monastersky 2005]，[Ewing 2006]，[Adler 2007]，[Hall 2007] 参照.）

(i) インパクト･ファクターが平均値であるとするのは正確ではない．なぜなら，多くの学術誌は，滅多に引用されない投書･論説等の非実質的記事を掲載しているが，これらの記事はインパクト･ファクター式の分母には数えられない．しかしながら，これらの記事は稀であるとはいえ引用されることもあり，インパクト･ファクター式の分子には**数えられる**．従ってインパクト･ファクターは，正確には，論文当たりの平均被引用

回数ではない．そのような「非実質的」記事を掲載する学術誌の場合，このずれは無視できなくなる．数学を含む多くの分野では，このずれはごく僅かである．

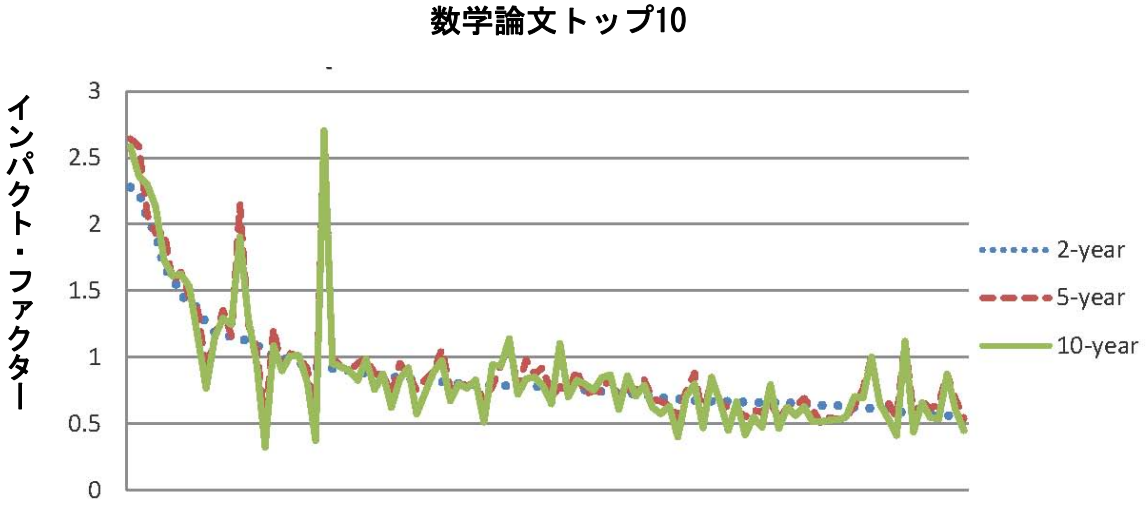
(ii) インパクト・ファクターの定義で使用する**2年の期間**は，統計を最新のものにすることを意図してのことである．[Garfield 2005]．生物・医科学等の分野では，殆どの論文が出版直後に引用されるため，この定義は適切である．しかし，数学等の他の分野の場合には，殆どの引用は2年の期間外で発生する．数学関係の学術誌における最近の300万件の引用データ(Mathematical Reviews引用データベース)を調べたところでは，学術誌の引用の約90%はこの2年の期間外に発生している．その結果、インパクト・ファクターは，被引用状況のたった10%程度に基づいており，被引用状況の大部分を見落としていることになる．(注4)



このグラフは，4分野で2003年に出版された学術誌における引用の年数を示す．2001-2002年に掲載された論文の引用が，インパクト・ファクターの計算に関与し，それ以外の年に掲載された論文はインパクト・ファクターの計算に無関係である．(トムソン・サイエンテフィック社のデータ)

それでは，2年間の期間の所為で，インパクト・ファクターは不当であろうか．数学関係の学術誌に関しては，証拠は曖昧である．トムソン・サイエンティフィック社の計算によれば，5年間インパクト・ファクターと通常の(2年間の)インパクト・ファクターとは良い相関関係にある由である [Garfield 1998]．Mathematical Reviews引用データベースを使って，数学関係学術誌の最もしばしば引用される上位100誌の2年，5年，10年の「インパクト・ファクター」(つまり，論文当たりの平均被引用回数)を計算できる．次のグラフで見れば，5年や10年の「インパクト・ファクター」が2年「インパクト・ファ

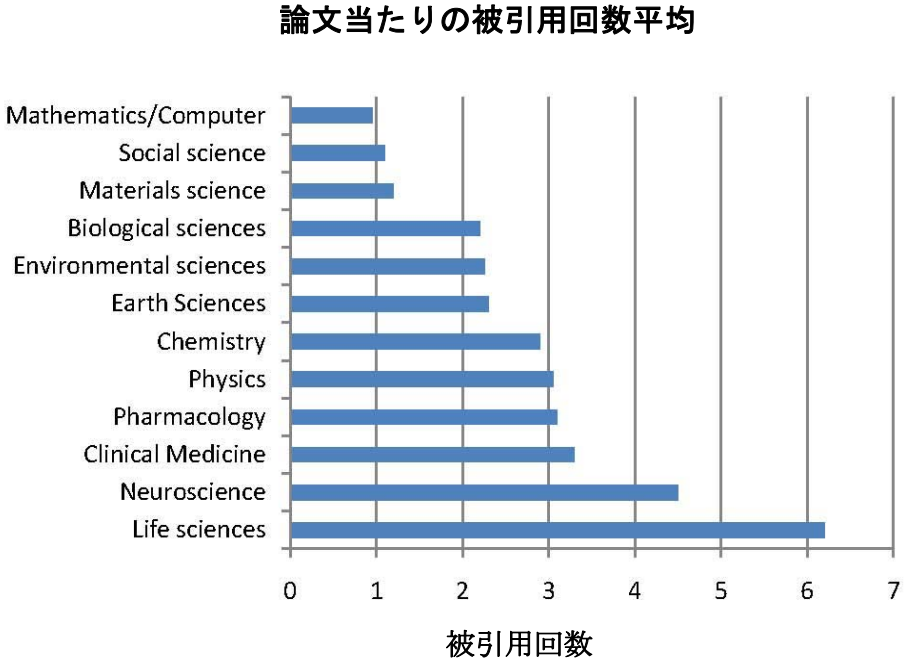
クター」にほぼ沿っていることが判る.



Mathematical Reviews引用データベースに基づく，数学学術誌100誌(訳注：横軸)に関する2年，5年，10年「インパクト・ファクター」

上に突出している外れ値の一つは，計算した期間中の一部で論文を公刊しなかった学術誌である．下の外れ値は掲載論文数の少ない雑誌である傾向があり，そのような学術誌に関するインパクト・ファクターの正常な変動性を反映しているに過ぎない．インパクト・ファクターの計算時に「ターゲット年」[訳注：被引用論文の発表年]を変更すれば，学術誌のランクが変わるのは明白であるが，インパクト・ファクターの小さな学術誌を除き，その変わり方は一般的に僅かである．インパクト・ファクターが小さい学術誌の場合，後述のように，「ソース年」[訳注：引用元論文の発表年]を変更しても，かなり変化する．

(iii) インパクト・ファクターは学術分野によりかなり異なる [Amin - Mabe 2000]．この違いの一部は上記の(ii) に起因する．即ち，ある分野における引用の多くが2年間の期間以外で発生するならインパクト・ファクターは遙かに小さくなる．他方で，違いの原因の一部は，単に，引用文化が分野により異なり，研究者は違った頻度や違った理由で引用しているからに過ぎない．(引用の意味は極めて重要なので，この点に関して後ほど詳述する.) 従って，異分野の学術誌を，インパクト・ファクターを使って意味ある形で比較するのは不可能である．



分野毎の論文当たり被引用回数平均. 引用の慣習が分野によりかなり異なることを示す. トムソン・サイエンテフィック社のデータに基づく. [Amin-Mabe 2000]

(iv) インパクト・ファクターは，年毎にかなり変動し，学術誌の規模が小さい程，その変動は大きくなる傾向がある［Amin - Mabe 2000］. 例えば，50論文より少なくしか掲載しない学術誌の場合，2002年から2003年へのインパクト・ファクターの変化はほぼ50%であった. もちろんのことながら，小規模学術誌のサンプル数が少ないので，当然期待通りである. しかし一方で，小規模学術誌のインパクト・ファクターの年毎の変動を無視して，特定年のインパクト・ファクターによって学術誌を比較してしまうのがしばしばである.

(v) 英語以外の言語で書かれた論文を掲載する学術誌の被引用回数は少なくなる傾向がある. なぜなら，大部分の研究者はそのような論文を読めない(あるいは読まない)からである. また，学術誌の質よりも種類のみが，インパクト・ファクターに影響を及ぼすことがあり得る. 例えば，書評を掲載する学術誌は，そうでないものよりも引用されることが遙かに多く，従ってインパクト・ファクターが(時として大幅に)高くなる［Amin - Mabe 2000］.

(vi) インパクト・ファクターに対する最も重要な批判は，その意味が良く理解されて

いないと言うことである．インパクト・ファクターを使って二つの学術誌を比較する場合，「より良い」と言うことが何を意味するかを定義する先験的モデルは存在しないのである．インパクト・ファクターが高ければ良い学術誌なのだと言う，インパクト・ファクター自身から派生するモデルしかないのである．古典的な統計パラダイムでは，まずモデルを定義し，（二つの学術誌に差がないとする）仮説を立て，統計量を求める．その値次第で，仮説を受け入れるか否かを決める．データそのものから情報（あるいはモデル）を導き出すことは，統計解析での正当なやり方であるが，この場合，どんな情報が導き出されたのかが明らかではない．インパクト・ファクターは質をどのように計るのか？質を計るのに最良の統計量か？それは一体何を**計っている**のか？（引用の意味するところに関する我々の後程の議論が，このことに関連する．）学術誌の質に関するモデルと，それのインパクト・ファクターとの関連に関しては，驚くほど何も判っていない．

インパクト・ファクターに関する以上の6つの批判は正しいが，インパクト・ファクターが大ざっぱなものであることを述べているに過ぎず，使い物にならないとは言っていない．例えば、インパクト・ファクターは，学術誌をグループ分けしてランク付けする出発点として使える．即ち，インパクト・ファクターを使ってまずグループ分けし，他の基準を使ってより精密なランク付けを行い，その上で最初のグループ分けが適切か否かを検証するのである．しかし，インパクト・ファクターを使用して学術誌を評価する際には注意を要する．例えば，異分野の学術誌を比較するのにインパクト・ファクターは使えない．また，インパクト・ファクターを使用する際には，学術誌のタイプの違いに細心の配慮が必要である．年毎の変化にも注意する必要がある．特に小規模学術誌では，小さな変動はランダムな現象に過ぎないことを理解する必要がある．また，学術誌すべてがデータ収録誌であるとは限らず，計測期間も短かすぎるため，分野によっては，インパクト・ファクターが引用状況を必ずしも正しく反映していないことを認識することが重要である．より長期間にわたる，より多くの学術誌に基づく別の統計なら，質をより良く示すかもしれない．最後に，引用は学術誌を評価する単なる一方法に過ぎず，他の情報により補足するべきもの（本報告書の中心的メッセージ）である．

上記は、統計に基づいてランク付けするどんな場合にも注意すべき事柄である．特定の年のインパクト・ファクターに基づいて学術誌を無分別にランク付けするのは，統計の誤用である．トムソン・サイエンティフィック社の名誉のために付言すると，同社はこれと同意見であり，インパクト・ファクターをそんな目的に使用する人達に（婉曲にではあるが）注意を呼びかけている．

「学術誌の有用性を評価するに当たって，トムソン・サイエンティフィック社では，インパクト・ファクターのみに依拠することはありませんし，皆様にも謹んで頂きたい．被引用度に影響を及ぼす様々な現象(例えば，平均的論文に引用されている参考文献数の平均）に対して細心の注意を払うことなしには，インパクト・ファクターを使用するべきではありません．インパクト・ファクターは事情通によるピア・レビューと併用すべきです.」［THOMSON: IMPACT FACTOR］

残念なことに、この忠告は非常にしばしば無視されている.

論文のランク付け

学術誌のランク付けをする際、インパクト・ファクターや，引用データに基づく類似の統計量は誤用され得るが，より根源的で，しかも知らない内に進行する誤用が存在する.個別の論文，研究者，プログラム，学問分野でさえもインパクト・ファクターを使用して比較しようとする誤用である．国境や学問分野を跨ってますます深刻化している問題であり，最近の全国規模の研究評価により一層深刻になっている．ある意味では，これは決して新しい現象ではない．研究者は，他の研究者の出版論文リストに関する意見を求められる機会がしばしばあるが，「彼女は優れた学術誌に発表している」や「彼の論文の殆どは低レベルの学術誌にしか発表していない」のようなコメントを耳にする．これらのコメントは賢明な評価である．ある研究者が普段(あるいは終始一貫して)発表する学術誌の質は，その研究者の研究全般を評価する際に使える多くの基準の一つである．しかし，インパクト･ファクターの所為で，学術誌の質を，その学術誌の掲載された各論文(および各著者)の質にも転嫁する傾向が強まってしまった.

トムソン・サイエンティフィック社は，この傾向を暗黙の内に助長している.

「インパクトが活用される，最近の最も重要な機会は，おそらく学術評価のプロセスにおいてである．研究者の論文が掲載された学術誌の権威の概要を全体的に把握する手段としてインパクト･ファクターを使える.」［THOMSON: IMPACT FACTOR］

この推奨がどのように解釈されているかを示す例が世界中の数学者から次のように寄せられている.

例1．私の大学では,サイエンス・サイテーション・インデックス・コア・ジャ

ーナル(Science Citation Index Core Journals)を使用した学術誌の新しい分類法を最近導入した．学術誌はインパクト・ファクターのみに基づいて，三つのグループに分類される．トップ・リストの中には学術誌が30誌あるが，数学の学術誌はその中には含まれていない．二番目のリストには667誌あり，その中に数学の学術誌が21誌含まれている．トップ・リストの中の学術誌に発表されれば，研究助成が3倍になり，二番目のリストなら2倍になる．コアリストの学術誌に掲載された論文には15点，トムソン・サイエンティフィック社のデータ収録誌に掲載された論文には10点が授与される．昇進に必要な最低点が決まっている．

例2．我が国では，大学の常勤教員は6年毎に評価される．学術面すべてで成功する鍵は,継続して良い評価を得ることである．評価に際しての，経歴以外での最も重要な基準は,発表論文5件のランキングである．最近では，トムソン・サイエンティフィック社のリストのトップ3分の1の学術誌に発表すれば3点，次の3分の1の学術誌に発表すれば2点，下位3分の1であれば1点ポイントが与えられる．（リストはインパクト・ファクターを使って作成される.）

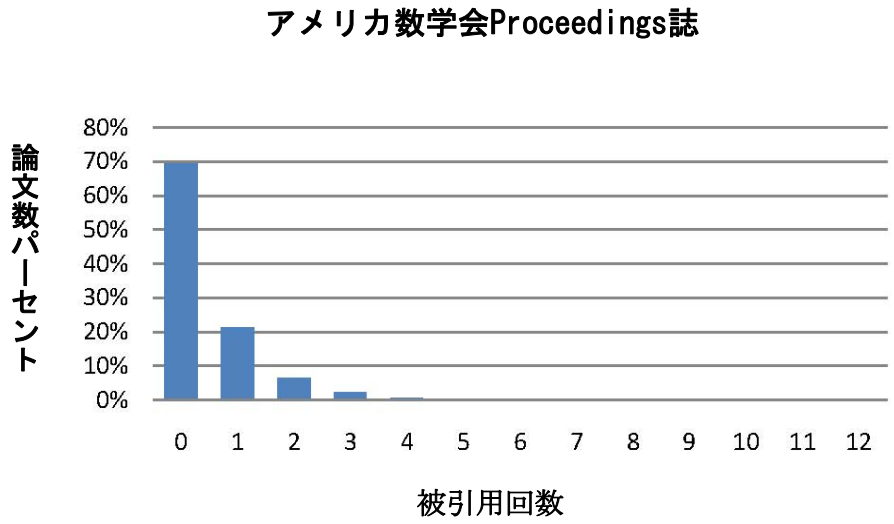
例3．我々の学科の教員は，単著相当換算論文数に掲載学術誌のインパクト・ファクターを掛けた数を含む公式によって評価される．昇進と雇用は，部分的にこの公式に基づく．

これらの例や我々に報告された数多の例では，個々の論文と著者を比較するのにインパクト・ファクターが明示的あるいは暗黙の内に使用されている．学術誌Aのインパクト・ファクターが学術誌Bよりも高ければ，当然のことながらAに掲載された論文はBに掲載された論文よりも優れており，従って，Aに掲載された論文の著者はBに掲載された論文の著者よりも優れていると評価するのである．この理屈は，学科や学術分野の比較にも，時として拡大解釈されている．

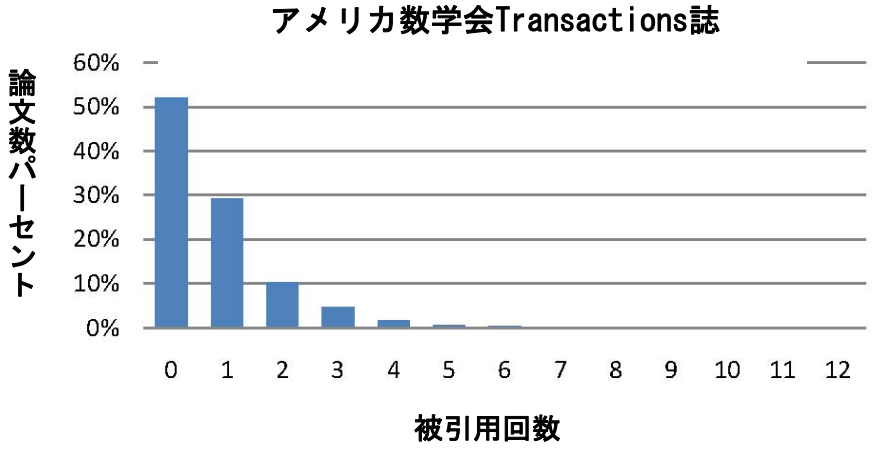
一つの学術誌に掲載された各論文の被引用回数の分布がかなり歪んでおり，ほぼいわゆる巾乗法則に従うことは，ずっと以前から知られている．([Seglen 1996], [Garfield 1987]）その結果どうなるかは，次の具体例で正確に示せる．

アメリカ数学会の会誌Proceedings of the American Mathematical Societyに2001年から2004年までの期間に掲載された論文の被引用回数の分布を次に示す．Proceedings誌は，通常10ページより短い論文を掲載する．当該期間中に掲載されたのは，2,381論文(約15,000ページ)である．Mathematical Reviews引用データベースに収録された2005

年発行学術誌を使用して計算してみると，Proceedings誌掲載論文の平均被引用回数(つまりインパクト・ファクター)は，0.434である．



アメリカ数学会の会誌Transactions of the American Mathematical Societyは，関連領域や内容がより実質的な長い論文を掲載する．上記と同一期間中に，Transactions誌は1,165論文(25,000ページ以上)を掲載し，掲載された論文の被引用回数は0から12までのばらつきがあった．平均被引用回数は0.846で，Proceedings誌のそれの約2倍であった．



ここで，二人の数学者がいて，一つの論文を一人はProceedings誌に，他方はTransactions誌に発表したとしてみよう．上述の例にある学科のやり方に従えば，2倍も高いインパクト・ファクターの学術誌に発表した二人目が一人目よりも優れていると評価されることになる！これは正当な評価であろうか？Transactions誌に掲載された論文は，Proceedings誌に掲載された論文よりも2倍優れているのであろうか?

Transactions誌に掲載された論文が，Proceedings誌に掲載された論文よりも(被引用回数に関して)優れていると主張するのであれば，平均ではなく，確率を問題にすべきなのである．間違っている確率はいくらか? つまり，無作為に抽出したProceedings誌掲載論文が，無作為に抽出したTransactions誌掲載論文以上の被引用回数を持つ確率はいくらか?

これは初歩的な計算で求められ，答えは62%である．つまり我々は62%の確率で間違っており，無作為に抽出したProceedings誌掲載論文は，無作為に抽出したTransactions誌掲載論文と同じくらい(あるいはより優れている)ことになる．Proceedings誌のインパクト・ファクターがTransactions誌の半分であるのに！ 正しいよりも間違っている方が高い確率なのである．多くの人はこの事実に驚くが，かなり歪んだ分布であることと，インパクト・ファクター計算の際に使用する期間が短い(全く引用されない論文が非常に多い理由)ことの当然の帰結なのである．(注5) 直感的観察よりも，正確な統計的考え方に価値のあることが判る．

これは学術誌の示す典型的な挙動であり，これら二つの学術誌を選んだことに特に意味はない．(例えば，Journal of the American Mathematical Society誌の同一期間に関するインパクト・ファクターは2.63であり，Proceedings誌の6倍である．しかしながら，被引用回数の意味では，無作為に抽出されたProceedings誌掲載論文は，32%の確率で，Journal誌掲載論文以上に良いことも判る．従って，ある学術誌のインパクト・ファクターが，それに掲載された論文に関して何らの情報も提供しないとするのは正しくないにしても，提供する情報は驚くほど曖昧であり，甚だしい誤解に繋がり得る．

個々の論文の被引用回数の代用に掲載誌のインパクト・ファクターを使用する上記3例のような計算は何ら合理的根拠を有しないことが判る．評価を行うに際して，半分以上(あるいは3分の1)の確率で正しくないことを主張するのは，決して良い方法でないことは明らかであろう．

個々の論文の被引用回数の代わりに掲載雑誌のインパクト・ファクターを代用するのが無意味であると判った以上，当該論文の著者，関与しているプログラム，そして関係する学術分野の評価にインパクト・ファクターを使用することは無意味であることになる．インパクト・ファクターや平均というものは，あまりにも曖昧であり，他の情報を援用

ある学術誌のインパクト・ファクターが，それに掲載された論文に関して何らの情報も提供しないとするのは正しくないにしても，提供する情報は驚くほど曖昧であり，甚だしい誤解に繋がり得る．

せずに正当な評価に使用するのは無理である.

個人をランク付けすることとその個人の発表論文をランク付けすることが別物であるのは当然である. しかし, その個人の論文を被引用回数のみを使ってランク付けしようというのなら, 当該論文の被引用回数を数え上げることから始めるべきである. 当該論文が掲載された学術誌のインパクト・ファクターは代替物として全く信頼できない.

研究者のランク付け

インパクト･ファクターは引用状況に基づく統計量としては最も有名であるが, 他の統計量が, 最近になって積極的に推奨されるようになった. 個人をランク付けするのに使用される統計量の例を三つ挙げてみよう.

h指数: 研究者S氏のh指数とは, 次のようなnの最大値である. 被引用回数がn以上であるS氏の論文の数がn編ある.

ここで例示する中で最も人気の高い統計量である. 研究者の被引用回数分布において, 右側の「裾」に焦点を当てることによって 「研究者の出力」を計る意図の下に, J. E. Hirsch [Hirsch 2006]により提案された. 発表論文数と被引用回数分布を単一数字で表すのが目的であった.

m指数: 研究者S氏のm指数とは, S氏のh指数を, S氏の処女論文以降の年数で割ったもの.

これも, 上記の論文中でHirschが提案したことである. 意図するところは, 論文を発表したり引用されたりする時間がまだ不十分な若手研究者に配慮するためである.

g指数: 研究者S氏のg指数とは, 次のようなnの最大値である. S氏の論文を被引用回数の多い順に並べたn番目までの論文の被引用回数の合計がn×n 以上.

これは2006年にLeo Egghe によって提案されたものである. h指数では, 被引用回数の多いn番目までの論文中に極めて被引用回数の多い論文のある可能性を考慮していないためである. g指数はそれに配慮している.

この他にも論文の経年や共著者数などを考慮した上記指数の変異形を含む様々な指標がある. ([Batista - Campiteli - Kinouchi - Martinez 2005], [Batista - Campiteli - Kinouchi 2006], [Sidiropouls - Katsaros - Manolopoulos 2006])

h指数を定義している論文中で,「研究者の研究成果の蓄積の持つ重要さ・意義・幅広いインパクトを容易に数えられる数としてh指数を提案した.」とHirsch は述べている.[Hirsch 2005, p. 5] 更に,「研究成果が重要な評価基準である場合に, 同一資源の獲得を目指して競い合っている複数の研究者を比較する際に役立つ.」と彼は付言している.

これらの主張のいずれも, 納得いく証拠によって裏付けられていない. h指数が, 研究者の研究成果の蓄積の重要さ・意義を計れるとの主張を裏付けるため, Hirsch はノーベル賞受賞者(別途, 科学アカデミー会員)グループのh指数を分析している. これらのグループに属する研究者のh指数は, 一般的に高いことを彼は示している. ノーベル賞受賞者であるが故に, これらの研究者のh指数が高いとも結論付けることができる.しかし, それ以上の情報なしにh指数が高いと言うだけでは, ある研究者がノーベル賞受賞者や科学アカデミー会員になれる確率は判らない. h指数の有効性を確立するためには, むしろそのような情報が必要なのである.

Hirshは, その論文中で, 二人の研究者を比較するためにh指数が使えるとも主張している.

> 「二人の研究者のh指数がほぼ同じなら, たとえ発表論文総数が違っても, あるいは被引用総数が違っても, 研究に関する二人の総合的なインパクトは比肩すると言いたい. 逆に、同じ研究年齢の二人の研究者の発表論文数と被引用回数がほぼ同じでも, h指数が大きく異なれば, h指数の高い方の研究者の方が熟達した研究者である可能性が高い.」[Hirsch 2005, p. 1]

これらの主張は, 常識で論破出来そうである. (例えば、二人の研究者が, 二人とも10回引用されている論文10編を持っているが, その内の一人が, 更に, 9回引用されている論文90編も持っているならどうであろうか. あるいは, 一人の研究者は10回引用されている論文を10編だけ持っているが, もう一人の研究者は100回引用されている論文を10編だけ持っているならどうであろうか. この二人が比肩すると誰が考えるであろうか) (注6)

Hirschはh指数の長所を絶賛して「研究者の研究出力を評価する単一数値による基準中で, h指数が他のものよりも好ましい・・・」と主張している. [Hirsch 2005, p. 1] しかし,「好ましい」の定義や, なぜ「単一数値による基準」を見つける**必要がある**のかという点に関しては何も説明していない.

このような方法に対する批判は多少存在するが，本格的な分析は殆どされていない．分析の大部分は，「妥当性の収束」即ち，h指数が，発表論文数や被引用総数等の公表・引用に関する計量と相関性が高いことを示しているにすぎない．これらすべては，研究公表という基本的現象の関数であり，示されている相関度は，取り立てる程のものではない．h指数に関する注目すべき論文[Lehmann - Jackson - Lautrup 2006]において著者達は，もっと注意深い分析を行い，h指数(実はm指数)は，単なる論文当たりの平均被引用回数程も「良く」はないことを示している．しかしこの場合でも，「良い」とは何を意味するかの説明が十分ではない．[Lehmann - Jackson - Lautrup 2006]のように古典的な統計パラダイムを適用すると，h指数は他の計量よりも信頼性が薄いことになってしまう．

どんな二人の研究者でも比較出来ることよりも，研究を理解することが目的であるべきである．

単一の学術分野内だけではなく，学術分野を跨って研究者の質を比較するためのh指数の変位形もいくつか考案されてきた．([Batista - Campiteli - Kinouchi 2006], [Molinari - Molinari 2008]) h指数を研究機関や学科の比較に使えると主張するものもいる [Kinney 2007]．これらは，複雑な被引用状況を単一の数値で捉えようとする，驚くほどナイーブな試みであることがしばしばである．確かに，被引用回数の単純なヒストグラムと比較した場合の，これら新しい指数の主要な利点と言えば，被引用状況に関する詳細の殆どを捨て去ってしまった結果，どんな研究者2名でもランク付けできるようになったことである．しかし，研究履歴を理解する上で，捨て去ってしまった情報が必要であることは，簡単な例でも判る．研究を評価する場合，どんな二人の研究者でも比較出来ることよりも，研究を理解することが目的であることは疑いない．

全国規模の評価機関が，h指数やその変形型の一つをデータの一部として収集していることがあるが，データの誤用である．残念ながら，単一数値で研究者をランク付けできることは極めて魅惑的であり，もっと単純な状況にあっても統計的推論の正しい用法をしばしば誤解してしまう一般の人々に拡がる恐れがある．

引用の意味

引用統計が研究の質を計る主要な方法であると推奨する人々は，引用が何を意味するか? という本質的な質問に答えてくれない．彼らは，引用回数に関する膨大なデータを収集し，統計量を導出するためにデータ処理し，その結果としての評価処理が「客観的である」と主張する．しかし，評価に繋がるのは，統計の**解釈**であり，その解釈は，引用の**意味**という極めて主観的なものに依拠している．

この手法を推奨する文献中に，引用の意味に関する明確な記述を見つけるのは驚くほど難しい.

> 「被引用指数の背後にある概念は，基本的に単純である．情報の価値は、それを使う人達が決めるということを認めれば，コミュニティー全体に与えるインパクトを計るのに勝る質の測定法はあるだろうか．学術コミュニティーの中で最も広範囲な集団(即ち,情報となる資料を使用もしくは引用する人達)が，我々の一連の知識に当該アイデアとその考案者が及ぼした影響やインパクトを決めるのである.」［THOMSON: HISTORY］

> 「各研究者の質を定量化するのは難しいが，論文をより多く発表している方が良く，論文の被引用回数(その分野における引用の慣習に依存して相対的に)は,質を計る際に役立つ量であるというのが一般的な見方である.」［Lehman - Jackson - Lautrup 2006, p. 1003］

> 「被引用頻度は，その学術誌がどれ程価値があり，どれ程使われているかを反映している・・・」［Garfield 1972, p. 535］

> 「内科医や生物医学の研究者が学術誌の論文を引用すると，その引用された学術誌が彼または彼女に何らかの影響を及ぼしたことを示している.」[Garfield 1987, p. 7］

> 「引用は，知的恩義を認めることである.」［THOMSON: FIFTY YEARS］

ここで関連している用語は「質」，「価値」，「影響」，「知的恩義」である．「インパクト」という用語は，Eugene Garfieldが引用指数(citation index)の作成という考えを推奨する短い論文を書いた1955年に初めて登場した用語であるが，今では引用を意味付ける普通名称となった．彼は次のように述べている.

> 「従って，重要性の極めて高い論文の場合，引用指数は定量的価値を持つ．なぜなら,当該論文の影響，つまりインパクト・ファクター，を歴史家が計る際に役立つからである.」［Garfield 1955, p. 3］

ここでも他と同様に,「インパクト・ファクター」と言う用語は，次のことを示唆するように意図されていることがかなり明白である．つまり，引用している論文は，被引

用論文の業績の「上に積み上げられた」ものであり，引用という仕組みによって研究が伝搬・前進することである.

引用の実際の意味するところは，上記の曖昧な説明が我々に信じ込ませようとしているよりもっと複雑であることを説く文献が沢山存在する．例えば，MartinとIrvineは，研究評価に関する1983年の論文で次のように述べている.

> 「質を計る際に引用を使うことが孕むこれら種々の問題点の根底にあるのは，著者達が特定の論文を引用し，他を引用しない**理由**に関する無知である．上記の問題は・・・単純な引用分析は，参考文献の挙げ方に関する極めて理性的なモデルを前提としている．つまり，引用する行為は，質が高いあるいは重要な過去の業績を科学的に好意に評価していることを主として反映するものとされており，今後引用しそうな著者達も，その特定の論文を引用する機会を等しく持っている・・・」［Martin－Irvine 1983, p. 69］

引用の意味するところは単純ではなく，引用に基づく統計は支持者が主張するほど「客観的」ではない.

Cozzensは，引用の意味に関する1988年の論文［Cozzens 1989］で，「引用行為は，学術的出版の根底にある，『褒美』と『修辞』という二つのシステムの結果である．」と主張している．一つ目は，引用と最もしばしば関連づけられる意味であり，引用している論文が，引用された論文に『知的恩義』を負っていることへの謝辞である．しかし，二つ目の意味するところはかなり違っており，引用された著者の得た成果ではおそらくない結果を説明している論文として参考に挙げていることである．このような修辞的な引用は，知的恩義を表明するのではなく，単に学術的会話を進行させているに過ぎない.

Cozzensの見たところでは，**殆どの**引用は修辞的である．このことは，殆どの現役数学者の経験でも確かめられる．（例えば，Mathematical Reviews引用データベースで，300万件以上の引用の内，ほぼ30%が論文ではなく著書の引用である.）このことがなぜ重要なのだろうか? 影響力・将来性のある論文を引用する「褒美」の場合と違って，修辞的引用には，種々の要因がある．つまり引用される著者の名声（「後光」効果）や，引用する側と引用される側の関係，学術誌の入手可能性(オープン・アクセス学術誌は，より被引用頻度が高いであろうか?)や，単一論文でいくつかの結果を参照できる利便性等である．これらの要因は，引用される論文の「質」と直接には殆ど関係していない.

「褒美」引用の場合でさえ，「流布性，貢献否定，作戦情報，説得力，貢献，読者への

警告，社会的合意」を含む様々な動機を反映している［Brooks 1996］．殆どの場合，これらの複数の動機で引用が行われる．著名な研究成果であっても，「抹消」効果の被害を被ることがある．つまり，他の研究者の研究成果に組み込まれてしまい，組み込んだ方がその後の引用先となってしまうのである．また，引用と言っても，顕著な研究に対する御褒美ではなく，不備な結果や推論に対する警告の対象として引用されることもある．本報告書では，そのような「警告」引用の例を多数紹介する．

引用の社会学は複雑な題材であり，本報告書で扱える範囲を超えている．しかし，これまでの大雑把な議論でも，引用の意味するところは単純ではなく，引用に基づく統計は支持者の主張するほど「客観的」ではないことが判る．

引用に基づく統計は，研究の質を評価する(例えばピア・レビューのような)他の方法と高い相関関係を有しており，引用の意味が何であろうと無関係だとする論者もいる．例えば，前述のエビデンス社報告では，高い相関性を理由に，他の評価法の代わりに引用統計が使える(使うべきである)と次のように論じている．

> 「エビデンス社は，文献計量的手法を使えば，研究者の認識と一致するような形で，研究の質を示す指標を作れるとしている．」［Evidence Report 2007, p. 9］

他の評価法と同じ評価結果が得られるので，それらを使うのを止めて，引用に基づく統計を使うべきであるとの結論のようであるが，循環論法であること以外にも，この論法の欠陥は明らかである．

統計の賢明な使用

研究評価に際して客観的な量(統計)に過度に依拠しようとするのは，何も目新しいことではなく，孤立した現象でもない．2001年出版の通俗書「とんでもないウソと統計(Damned lies and statistics)」［訳注：和訳タイトルは「統計はこうしてウソをつく」］で，社会学者Joel Bestは説得力ある形で次のように述べている．

> 「ある種の物には神秘な力が宿っていると信じる文化がある．人類学者はそのようなものを物神(fetish)呼ぶ．我々の社会で，統計は一種の物神である．統計を不思議な力を持つ，単なる数字以上のものと見なす傾向がある．統計は，真実の力強い表現であり，現実世界の複雑さと混乱状態から単純な真理を抽出するかのように，我々は振る舞っている．我々は統計を使って，社会の複雑な問題をより判りやすい推定値・割合・比率に変換する．統計は我々の関心を方

向付け，我々が何に関してどの程度心配すべきかを示してくれる．ある意味で，社会的な問題が統計に様変わりしてしまう．統計のことを，真実で議論の余地のないものと我々がみなすため，社会問題の見方を支配する物神のような不思議な力を統計が獲得してしまう．我々は統計のことを，我々が作り出す数字ではなく，発見する真理と考えてしまう.」[Best 2001, p. 160]

引用統計の持つ不思議な力に対する信仰は，全国レベルあるいは研究機関レベルの研究評価活動関係の文書の至る所に見出すことができる．またh指数や派生型の推奨者達の著作中にも見つけることができる.

ページ･ランク付けアルゴリズムのように，より高度な数学的アルゴリズムなどを使用した引用分析によりインパクト・ファクターを改良しようという最近の試みの中でも，このような態度は歴然としている．([Bergstrom 2007], [Stringer - Sales - Pardo - Nunes 2008]) 改良版の推奨者達は，その効能をいろいろ謳っているが，分析に基づいた正当化もしておらず，評価するのは困難である．しかも，いっそう複雑な計算に基づいているため,その背後にある(しばしば隠された)仮定を識別するのは，殆どの人にとって容易ではない．(注7) 我々は数値とランク付けを，我々の作り出したものとしてではなく，畏敬の念を持って真実として扱わなければならないとされている.

薬を使う際には医者に相談するように，統計を使う際には統計学者に相談すべきである.

公的資金による活動で精査の対象となったのは，研究活動が初めてではなく，教育システム(学校)から医療(病院や個々の外科医までも)の活動実績を定量的に評価しようする試みが過去数十年にわたって行われてきた．いくつかの場合には，統計専門家達が乗り出し,賢明な計量や統計の適切な使い方を計測担当者達に助言した．薬を使う際には医者に相談するように，統計を使う際には統計学者に相談すべきである．[Bird 2005] と [Goldstein - Spiegelhalter 1996]に素晴らしい例が二つ挙げられている．これらの2例は，前者が公営企業の実績監視，後者が医療や教育と，研究以外の実績評価を扱っているが，どちらの例も，研究評価の際の賢明な統計活用法に関して洞察を提供してくれる.

特にGoldsteinとSpiegelhalterの論文は，単純な数字(例えば，生徒の成績や医学的結果)に基づく成績表（ランク付け)を取り扱っており，引用統計を使って，学術誌・論文･研究者をランク付けすることによって研究評価を行うことと特に関連する．彼等はその論文で，どんな実践評価にも通用する，次のような三要素で構成される枠組を概説している.

データ

> 「どんなにうまく統計を操っても，収集されたデータが不適切であったり，不正直であったりすることを克服することはできない.」[Goldstein - Spiegelhalter 1996, p. 389]

これは引用に基づく実績評価にとって重要な所見である．例えばインパクト・ファクターの場合は，トムソン・サイエンティフィック社が選定したデータ収録誌のみから得られた，データの部分集合に基づいている．（インパクト・ファクター自身が，選定に際しての主要な基準であることに留意する.）このデータの完全性(整合性)に対して疑問を呈する人達もいる．[Rossner - VanEpps - Hill 2007]．また，他のデータ集合の方がもっと完全性が高いと指摘する人達もいる． [Meho - Yang 2007] h指数のように，引用統計を実施するために，Google Scholarを使おうとするグループもある．しかし，Google Scholarに含まれているデータは(著者の氏名がウェブ登録から自動的に抽出されるため)しばしば不正確である．著者が一意的に識別できないために，個々の研究者の引用統計を得るのが難しい場合がある．そのため，状況や国次第では，正確な引用データを集めるのに大きな障害になり得る．引用分析にどんなデータ集合を使うかはしばしば見落とされる．欠陥のあるデータに基づいた統計からは，誤った結論を出してしまいそうである．

統計分析と提示

> 「適切な統計**モデル**の設定，全結果の提示において結果に含まれる**不確実性**の決定的な重要性，交絡因子を考慮した結果の**調整**の手法，そして最後に明示的な**ランク付け**がどれ程信頼するに足るか，に我々は特に注目する.」[Goldstein - Spiegelhalter 1996, p. 390]

前述したように，論文・研究者・プログラムのランク付けに引用統計が使用される殆どの場合，何らのモデルも事前に特定されない．むしろ，データそのものが，しばしば曖昧なモデルを特定している．この循環論法では，ある対象が(データベース中で)ランクが高い故に，高いランクが付けられることになる．**どの**ランク付けでも，不確定性に関しては殆ど注意が払われず，また，その不確実性(例えばインパクト・ファクターの年次的変動)がランク付けにどう影響するかに関して,殆ど分析していないのがしばしばである．最後に，交絡因子(例えば，特定の学術分野，学術誌が掲載する論文のタイプ，研究者が実験屋か理論屋か)は，ランク付けに際してしばしば，特に全国規模の実績評価の際には，無視される．

解釈とインパクト

「本論文で論じる比較は公益性が極めて高いが，その限界に注意を払うことが不可欠であるにもかかわらず無視されそうである．調整された結果が研究機関の『質』に関する妥当な計測であるかも問題であるし，また分析者は，研究機関や研究者が『ランク』を上げるために行う行動変更が，結果にどのような影響をもたらすかにも注意する必要がある．」［Goldstein－Spiegelhalter 1996，p. 390］

研究の評価は大きな公益性も有している．個々の研究者にとっては，経歴に深刻で長期的な影響を及ぼす．学科にとっては，ずっと将来に至る成功の見通しを変えてしまう．ある学術分野にとっては，評価次第で停滞するか活況を呈するかが決まってしまう．そのように重要な作業を行うからには，それを実施する際のツールの妥当性と限界を理解するべきである．引用は研究の質をどの程度計ることができるであろうか? 被引用回数は質と相関しているように見えるし，質の高い論文はよりしばしば引用されると直感的に理解できる．しかし，前述したように，特に学術分野のいくつかでは，質の高さ以外の理由で頻繁に引用される論文が存在し，従って，頻繁に引用されることが必ずしも高い質を意味しない．引用統計に基づくランク付けの正確な解釈を，もっと良く理解する必要がある．更に，もし引用統計が研究評価において中心的役割を演ずるのなら，著者，編集者，更には出版者さえもが，自分達に有利に働くようにシステムを操る方法を見付け出すであろう［Macdonald－Kam 2007］．このことが長期的にどのような影響を及ぼすかは，はっきりしないし，検討されてもいない．

GoldsteinとSpiegelhalterによる論文は，研究評価に際して無邪気な統計に過度に依拠することが決して孤立した問題ではないことをはっきりと指摘しており，今日でも読む価値がある．政府・研究機関・研究者は，異なった文脈において過去にも似たような問題と取り組み，統計ツールをよりよく理解して他の評価手段で補足する方法を見出してきた．GoldsteinとSpiegelhalterは前向きな希望の表明で論文を締めくくっている．

「最後に，我々は機関評価の数多くの試みに対し批判的であったが，そのような比較すべてが必然的に不備であると信じているとの印象を与えたくない．機関を比較することや，機関がなぜ互いに異なるのかを理解することは極めて重要な作業であり，対決よりも協調の精神に基づいて実施されるのが最善のようである．おそらくそれが，理解と改善に繋がるような客観性に基づく情報を獲得できる唯一の確実な方法であろう．**我々が批判しようとした，極度に単純化した手法の本当の問題点は，このようにより立派な目的から注意と資源をそらしてしまうことである．**」［Goldstein－Spiegelhalter 1996，p. 406］

研究評価に関与する誰もが共有するべき目標の中で, 上記より優れたものを見付けるのは困難であろう.

定量的研究評価に関する
国際数学連合(IMU),
応用数理国際評議会(ICIAM),
数理統計学会(IMS)
合同委員会報告

ロバート・アドラー(Robert Adler, イスラエル工科大学テクニオン),
(委員長)ジョン・ユーイング(John Ewing, アメリカ数学会),
ピータ・テ―ラー(Peter Taylor, メルボルン大学)

参考文献

Adler, Robert. 2007. The impact of impact factors. *IMS Bulletin*, Vol. 36, No. 5, p. 4.
http://bulletin.imstat.org/pdf/36/5

Amin, M.; Mabe, M. 2000. Impact factor: use and abuse. *Perspectives in Publishing*, No. 1, October, pp. 1--6.
http://www.elsevier.com/framework_editors/pdfs/Perspectives1.pdf

Batista, Pablo Diniz; Campiteli,Monica Guimaraes; Kinouchi, Osame; Martinez, Alexandre Souto. 2005. Universal behavior of a research productivity index. arXiv: physics, v1, pp. 1--5.
arXiv:physics/0510142v1

Batista, Pablo Diniz; Campiteli,Monica Guimaraes; Kinouchi, Osame;. 2006. Is it possible to compare researchers with different scientific interests?. *Scientometrics*, Vol 68, No 1, pp. 179--189.
http://dx.doi.org/10.1007/s11192-006-0090-4

Bergstrom, Carl. Eigenfactor: measuring the value and presitige of scholarly journals. College & Research Libraries News, Vol 68, No. 5, May 2007
http://www.ala.org/ala/acrl/acrlpubs/crlnews/backissues2007/may07/eigenfactor.cfm
(See also **http://www.eigenfactor.org/methods.pdf**.)

Best, Joel. 2001. Damned lies and statistics: untangling the numbers from the media, politicians, and activists. University of California Press, Berkeley.
(ジョエル・ベスト著, 林 大訳「統計はこうしてウソをつく---だまされないための統計学入門」白楊社, 2002)

Bird, Sheila; et al. 2005. Performance indicators: good, bad, and ugly; Report of a working party on performance monitoring in the public services. *J.R.Statist. Soc* A (2005), 168, Part 1, pp. 1--27.
http://dx.doi.org/10.1111/j.1467-985X.2004.00333.x

Brooks, Terrence. 1986. Evidence of complex citer motivations. *Journal of the American Society for Information Science*, Vol 37, No. 1, pp. 34--36, 1986.
http://dx.doi.org/10.1002/asi.4630370106

Carey, Alan L.; Cowling, Michael G.; Taylor, Peter G. 2007. Assessing research in the mathematical sciences. *Gazette of the Australian Math Society*, A.L. Carey, Vol. 34, No. 2, May, pp. 84--89.
http://www.austms.org.au/Publ/Gazette/2007/May07/084CommsCarey.pdf

Cozzens, Susan E. 1989. What do citations count? The rhetoric-first model. *Scientometrics*, Vol 15, Nos 5--6, (1989), pp. 437--447.
http://dx.doi.org/10.1007/BF02017064

Egghe, Leo. 2006. Theory and practice of the g-index. *Scientometrics*, vol. 69, No 1, pp. 131--152.
http://dx.doi.org/10.1007/s11192-006-0144-7

Evidence Report. 2007. The use of bibliometrics to measure research quality in the UK higher education system. (大学における研究政策に関する英国の委員会(the Research Policy Committee of Universities, UK)の報告書をエビデンス社が製作したもの．エビデンス社は，研究実績分析と解釈を専門とする会社であり，トムソン・サイエンティフィック社と「戦略的」同盟関係にある.)
http://bookshop.universitiesuk.ac.uk/downloads/bibliometrics.pdf

Ewing, John. 2006. Measuring journals. *Notices of the AMS*, vol. 53, no. 9, pp. 1049--1053.
http://www.ams.org/notices/200609/comm-ewing.pdf

Garfield, Eugene. 1955. Citation indexes for science: A new dimension in documentation through association of ideas. *Science*, 122(3159), p.108--11, July 1955.
http://garfield.library.upenn.edu/papers/science1955.pdf

_____. 1972. Citation analysis as a tool in journal evaluation. *Science*, 178 (4060), pp. 471--479, 1972.
http://www.garfield.library.upenn.edu/essays/V1p527y1962-73.pdf

_____. 1987. Why are the impacts of the leading medical journals so similar and yet so different? *Current Comments* #2, p. 3, January 12, 1987.
http://www.garfield.library.upenn.edu/essays/v10p007y1987.pdf

_____. 1998. Long-term vs. short-term journal impact (part II). *The Scientist* 12(14):12-3 (July 6, 1998).
http://garfield.library.upenn.edu/commentaries/tsv12(14)p12y19980706.pdf

_____. 2005. Agony and the ecstasy—the history and meaning of the journal impact factor. Presented at the *International Congress on Peer Review and Bibliomedical Publication*, Chicago, September 16, 2005.
http://garfield.library.upenn.edu/papers/jifchicago2005.pdf

Goldstein, Harvey; Spiegelhalter, David J. 1996. League tables and their limitations: Statistical issues in comparisons of institutional performance. *J R. Statist. Soc. A*, 159, No. 3. (1996), pp 385--443.
http://links.jstor.org/sici?sici=0964-1998%281996%29159%3A3%3C385%3ALTATLS%3E2.0.CO%3B2-5
http://dx.doi.org/10.2307/2983325

Hall, Peter. 2007. Measuring research performance in the mathematical sciences in Australian universities. *The Australian Mathematical Society Gazette*, Vol. 34, No. 1, pp. 26--30.
http://www.austms.org.au/Publ/Gazette/2007/Mar07/26HallMeasuring.pdf

Hirsch, J. E. 2006. An index to quantify an individual's scientific research output. *Proc Natl Acad Sc USA*, Vol. 102, No. 46, pp. 16569--16573.
http://dx.doi.org/10.1073/pnas.0507655102

Kinney, A. L. 2007. National scientific facilities and their science impact on nonbiomedical research. *Proc Natl Acad Sci USA*, Vol. 104, No. 46, pp. 17943--17947.
http://dx.doi.org/10.1073/pnas.0704416104

Lehmann, Sune; Jackson, Andrew D.; Lautrup, Benny E. 2006. Measures for measures, *Nature*, Vol 444, No. 21, pp. 1003--1004.
http://www.nature.com/nature/journal/v444/n7122/full/4441003a.html

Macdonald, Stuart; Kam, Jacqueline. 2007. Aardvark et al.: quality journals and gamesmanship in management studies. *Journal of Information Science*, Vol. 33, pp. 702--717.
http://dx.doi.org/10.1177/0165551507077419

Martin, Ben R. 1996. The use of multiple indicators in the assessment of basic research, *Scientometrics*, Vol 36, No. 3 (1996), pp. 343--362.
http://dx.doi.org/10.1007/BF02129599

Martin, Ben R., Irvine, John. 1983. Assessing basic research. *Research Policy*, Vol 12 (1983), pp. 61--90.
http://dx.doi.org/10.1016/0048-7333(83)90005-7

Meho,Lokman; Yang, Kiduk. 2007. Impact of data sources on citation counts and rankings of LIS faculty: Web of Science vs. Scopus and Google Scholar. *Journal of the American Society for Information Science and Technology*, Vol 58, No 13, pp. 2105--2125.
http://dx.doi.org/10.1002/asi.20677

Molinari, J. F., Molinari, A. 2008. A new methodology for ranking scientific institutions. To appear in *Scientometrics*
http://imechanica.org/files/paper.pdf

Monastersky, R. 2005. The number that's devouring science. *Chronicle Higher Ed.* Vol. 52, No. 8.
http://chronicle.com/free/v52/i08/08a01201.htm

Rossner, Mike; Van Epps, Heather; Hill, Emma. 2007. Show me the data. *Journal of Cell Biology*, Vol 179, No 6, December 17, pp. 1091--1092.
http://dx.doi.org/10.1083/jcb.200711140

Seglen, P. O. 1997. Why the impact factor for journals should not be used for evaluating research; *BMJ*, 314:497 (15 February).
http://www.bmj.com/cgi/content/full/314/7079/497

Sidiropoulos, Antonis; Katsaros, Dimitrios; Manolopoulos, Yannis. 2006. Generalized h-index for disclosing latent facts in citation networks. V1, arXiv:cs.
arXiv:cs/0607066v1 [cs.DL]

Stringer MJ, Sales-Pardo M, Nunes Amaral LA (2008) Effectiveness of journal ranking schemes as a tool for locating information. PLoS ONE 3(2): e1683
http://dx.doi.org/10.1371/journal.pone.0001683

THOMSON: JOURNAL CITATION REPORTS. 2007. (Thomson Scientific website)
http://scientific.thomson.com/products/jcr/

THOMSON: SELECTION. 2007. (Thomson Scientific website)
http://scientific.thomson.com/free/essays/selectionofmaterial/journalselection/

THOMSON: IMPACT FACTOR (Thomson Scientific website)
http://scientific.thomson.com/free/essays/journalcitationreports/impactfactor/

THOMSON: HISTORY (Thomson Scientific website)
http://scientific.thomson.com/free/essays/citationindexing/history/

THOMSON: FIFTY YEARS (Thomson Scientific website)
http://scientific.thomson.com/free/essays/citationindexing/50y-citationindexing/

注記

(注1) 1977年10月のリーダース・ダイジェスト誌は，本引用文をアインシュタインによるものとしており，次のような実際の発言が根拠となっているようである．「経験から得られたデータのいずれもを適切に表現できることはあくまでも譲らずにいながら，それ以上還元できない基礎的要素を最も単純かつ最も少なくするのが，あらゆる理論の至上目的であることは否定できない．」 オックスフォード大学Herbert Spencer講演「理論物理の方法について」(1933年6月10日)；Philosophy of Science 第1巻2号（1934年4月)，pp.163 - 169にも収録.

(注2) 本節ではトムソン・サイエンティフィック社のインパクト・ファクターに焦点を当ててきたが，同社はそれ以外の二つの統計量の使用も推奨していることに留意する．また，学術誌の平均被引用回数に基づく同様の統計は，Scopus，Spires，Google Scholar，(数学では) Mathematical Reviews引用データベースを含む他のデータベースからも導出することができる．Mathematical Reviews引用データベースは，2000年から現在までの400以上の数学関係学術誌による300万件以上の文献引用データであり，引用対象はMathematical Reviews誌が採り上げた1940年以来の文献である.

(注3) トムソン・サイエンティフィック社の2008年3月発表では，次のカテゴリーでデータを収録誌しており，重複もあって総計は400学術誌である.

数学（217）
応用数学（177）
学際的数学（76）
数理物理（44）
確率・統計（96）

それとは対照的に，Mathematical Reviews誌は，毎年1,200誌を遙かに超える学術誌を採り上げる対象としており，その内800誌以上を「コア」（掲載論文の全てをMathematical Reviews誌で採り上げると言う意味で)と考える．Zentralblatt誌もほぼ同数の数学学術誌を採り上げている.

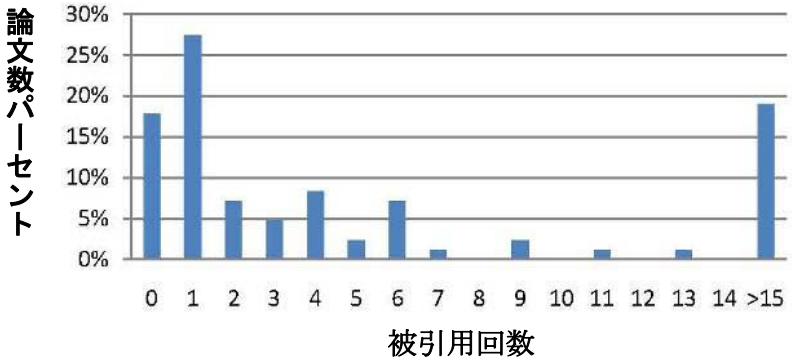
(注4) Mathematical Reviews引用データベースは，(2008年3月現在)約400学術誌の掲載論文が2000年から現在までに引用した文献を300万件以上収録している．これらの被引用文献は，Mathematical Reviewsデータベース収録の文献と照合されており，数十年以上も遡る．Science Citation Indexの場合と異なり，学術誌のみではなく単行本の引用データも含む．興味深いのは，引用の約50%は過去10年間の文献であり，25%はその前の10年間，12.5%は更にその前の10年間等々と言うことである．このような傾向は，学

術分野に依るのはもちろんである.

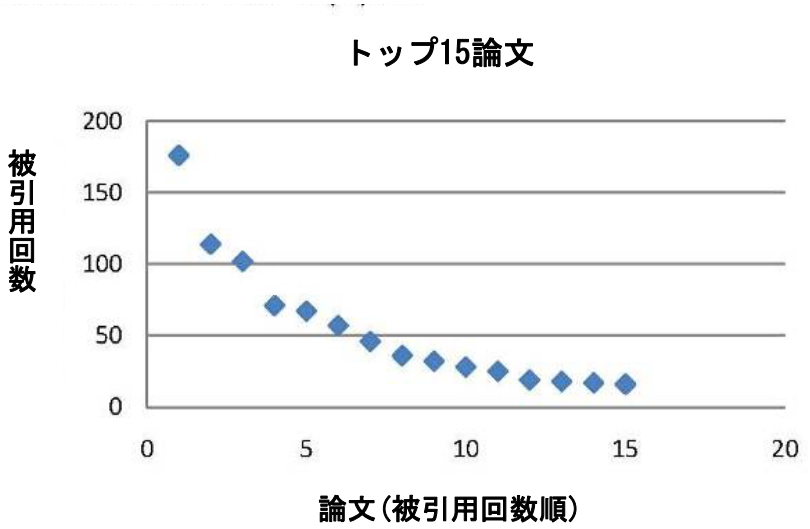
(注5) 歪んだ分布と狭い窓(引用元としては単年度の学術誌, 引用先は5年間)の結果, 殆どの論文は全く引用されないかあるいは引用されたとしてもごく僅かとなる. 無作為に選んだ2論文の引用状況がほぼ同じになってしまうのは, 直感的に明らかであろう. 多くの論文が全く引用されないかあるいは引用されたとしてもごく僅かになってしまうのは, 数学において引用されるまでの経過時間の長さ(何年もかかる)も原因である. 引用元となる学術誌の「ソース年」と引用先となる学術誌の「ターゲット年」の両方の期間をより長く設定すれば, 被引用回数はかなり増加し, 被引用状況によって学術誌を評価するのがより容易になるであろう. これは引用分析に際して[Stringer - et al. 2008]が使用した方法である. 従って, もし十分長い期間を設定すれば, 個々の論文の被引用回数の分布は対数正規であるであることが判る. 従って, 二つの学術誌を被引用分布により比較することが考えられ, インパクト・ファクターを使用するよりも高度であることは間違いない. しかし, いずれにせよ被引用状況のみしか考慮していないのには変わりない.

(注6) h指数のみを使用すると, どれ程の情報を失うかを示すために, 実在する一流の中堅数学者で84編の論文を発表済みの方の例を挙げてみよう. 被引用状況の分布は次のようである.

ある研究者の被引用データ(84論文)



20%を少し下回る論文[訳注: 以下では, 約17.8%と考え15編と見なしている]が15回以上引用されていることに注目して頂きたい. これら15編の論文の被引用回数の分布は次のようになっている.



しかし，Hirschの分析では，ｈ指数が15であるということ，即ち上位15論文が15回以上引用されていること，以外の情報は全て捨て去られてしまう．

(注7) [Bergstrom 2007]にあるアルゴリズムは，各引用にウエイトを付けるページ・ランク・アルゴリズムを使用し，被引用の加重平均を使用した「インパクト・ファクター」を計算する．ページ・ランク・アルゴリズムは，引用の「価値」を考慮するという利点がある．その一方で，最終結果を理解するのがより困難となるが故に，この方法の複雑さは危険を伴い得る．この場合は，すべての「自己引用」を捨て去る．つまり、学術誌Ｊに掲載された論文が、それ以前の5年間に同一学術誌Ｊに掲載された論文を引用しても考慮しないということである．これは「自己引用」の常識で考えた意味ではなかろう．Mathematic Reviews引用データベースのデータを一見しただけでも，引用の約3分の1を捨て去ることになることが判る．

[Stringer‐et al. 2008]にあるアルゴリズムは興味深い．引用の時間スケールの問題や，ある学術誌に掲載された無作為抽出論文と別の学術誌に掲載された無作為抽出論本を比較する問題を取り扱っているからである．この場合でも，アルゴリズムが複雑であるため，結果を評価するのは，大部分の人にとって困難である．論文の第2ページに，「学術誌Jに掲載された論文の『質』の分布は正規分布であると最初に仮定する･･･」との注目すべき仮定が忍び込んでいる．この仮定は，通常の経験とは異なるようである．