International journal of basic and applied research
www.pragatipublication.com
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

The Most-Cited Statistical Papers

Mrs. B. Padmaja
Department of Statistics
Samata Degree College

ABSTRACT: We make an effort to categorize statistical studies by citation count and provide short comments on each of the 25 papers that make the cut. The majority of the articles on this list address the multiple comparisons issue, non-parametric approaches, or their uses in the biological sciences. Also included are the most-cited articles from 1993 and following years. These articles mostly concern Bayesian approaches and wavelets, in contrast to the most-cited publications in general. Using citation counts raises a number of concerns, some of which are briefly addressed below.

KEY WORDS: Citations, history of statistics

Citations in General

The statistical literature has been mostly ignored in the many discussions on citation counts, with the notable exceptions being works by Stigler (1994), Altman & Goodman (1994), and Theoharakis & Skordia (2003). While Gilbert (1977) and Edge (1979) have looked at citation counts as indicators of the impact of research, Austin (1993) evaluated their dependability in academic tenure and promotion choices. Also refer to Cronin (1984).

Citation counts are being employed too much to assess intellectual literacies, according to Edge (1979). Some have voiced similar complaints. Despite these concerns, citation counts seem to be gaining popularity. One way in which PhD programs in statistics and related subjects are ranked is by their citation rates, according to the National Research Council. Not only are citation counts used to rank scientific publications, but they also seem to be employed more and more in academic advancement choices. For instance, according to ISI Journal Citation Reports, Statistical Science scored #5 in citation impact factor and #16 in total citations received in 2002 out of 71 journals in the Statistics and Probability category, with 1,051 citations.While researching what makes a paper popular with readers, Donoho (2002) offered advice on how to write citation-worthy pieces. "Develop a method which can be applied on

International journal of basic and applied research
www.pragatipublication.com
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

statistical data of a kind whose prevalence is growing rapidly" was his first priority. As an example, a study would definitely get a lot of citations if someone could invent "the" way to data mining.

It is not unexpected that "Implement the method in software, place examples of the software's use in the paper, make the software of broad functionality, and give the software away for free" was number two on Donoho's list, given that new methodology is typically implemented in software after a few years.

According to Garfield (1998), most scientific works referenced between 1945 and 1988 were cited only once. Hamilton (1990) found that out of all the articles published between 1981 and 1985, 55 percent got no citations within five years after publication. This finding is contentious. Hamilton (1991) used the same data to dissect the 55% of uncited publications and found a wide range of fields with varying percentages of uncited papers, from 86.9% in engineering to 9.2% in atomic, molecular, and chemical physics. In contrast to Hamilton's findings, Pendlebury (1991) found that only 22.4% of 1984-published scientific publications were still uncited in 1988.

Various areas have varying rates of paper citations. According to ScienceWatch (1999), papers in mathematics required a minimum of 291 citations to rank in the top 0.01% from 1981 to 1997, but papers in molecular biology and genetics required 1,823 citations for the same position.

How can we interpret these figures? These findings, a skeptic may argue, prove that most research is useless. Although there have been many cases of scholars failing to credit relevant work, it seems that the majority of published articles do not impact the work of other researchers. If we look at the distribution of publications according to their influence, we can observe that there is a significant degree of right skewness.

The 25 Most-Cited Papers

Our ranking of the 25 statistical articles with the highest number of citations is shown below. The main statistics journals were not our exclusive focus. Only articles in which a new statistical approach was either suggested, an existing method was revised, or an existing method was used in a creative manner to solve a significant scientific issue were evaluated for inclusion. Because of the subjective nature of the criteria, their application is certain to be subjective. There is no agreed-upon definition of "statistics," as Straf (2003) points out.

International journal of basic and applied research
www.pragatipublication.com
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

According to the ISI Web of Science (as of 1 December 2003), these are the citation counts. Since not all scientific publications are included in the Web of Science, the numbers are all undercounts. Furthermore, we made no effort to rectify inaccurate citation information, such as citations with the wrong page or volume numbers. Considering these parameters may cause a shift in the top 25 or possibly the removal of certain articles. But the biggest problem is that citations before 1945 aren't included in the ISI Web of Science citation counts. This creates an issue for publications that were published before that period. To make matters worse, the early studies had a harder difficulty getting citations since there weren't as many scientific journals back then. The citation rate of the original publication proposing a technique also falls when the approach becomes widely used in statistics (e.g., the one-sample t-test). Additionally, for some articles, we determined their current yearly citation rate, which is shown in parentheses where the data is available. Due to the short time frame (less than six months) in which this estimate was derived in late 2003, it is considered cautious.

Even though fuzzy logic and statistics are related, certain highly cited articles on the topic—for example, Zadeh's (1965) work, which received 5,022 citations, or 338 per year—were not classified as statistical publications (1995). Likewise, Hopfield (1982) was left out as well, despite having 3,574 citations (156 per year) on the subject of neural networks. Because the suggested technique was so simple, Reed & Muench (1938)—which received 10,974 citations, or 242 per year—was omitted. Due to the belief that the methodology and findings were mostly based on probability rather than statistics, Wright (1931) was left out, despite having 2,218 citations (144 per year). Furthermore, we did not think that the articles by famous statisticians Cooley and Tukey (1965) and Nelder and Mead (1965) had sufficient statistical support to be included in our list. For example, Cooley and Tukey received 2,872 citations, or 78 per year, whereas Nelder and Mead received 5,635 citations, or 426 per annum. Due to the inclusion of Cooley & Tukey (1965) in Kotz & Johnson (1997), which was introduced by I. J. Good, some of these judgments are open to debate.

It would be hardly unexpected if our investigation failed to capture some manuscripts that had substantial statistical information. Furthermore, it may be argued that a few of the papers included in our compilation ought to have been omitted for various reasons. Any feedback from our paper's audience on these matters is much appreciated.

The following is our list with some brief commentary.

(1) With 25,869 citations (currently cited 1,984 times per year),

Using centered survival data, Kaplan and Meier (1958) suggested a non-parametric approach to estimate the proportion of objects in a population whose lifespan surpassed a certain period t. In

Index in Cosmos
April 2019 Volume 9 ISSUE 2
UGC Approved Journal

medical research, this kind of information is commonplace. Not only does this work rank in the top five most referenced papers in the whole scientific area, but it also has the greatest amount of citations among statistics publications. The total number of citations earned by this work surpasses that of all papers published in the Journal of the American Statistical Association in 2002, according to data from Journal Citation Reports. An influential work in statistics, this article was published in Kotz & Johnson (1992b, pp. 311 - 338) and was introduced by N. E. Breslow.

According to Kaplan (1983), he and Meier really submitted different manuscripts to the Journal of the American Statistical Association. Their similarities led the editor to suggest merging the two articles into a single document. They worried that someone else may publish their concept for the whole four years it took them to work out their differences in approach.

The interesting thing is that this article was used by Garfield (1989) to illustrate how a piece of work might take a while to be noticed. It is worth noting that the study did not garner many citations annually until the early 1970s, namely within the first fifteen years after its publication (as seen in Figure 3 of Garfield (1989)). From 1958 to 1968, it was referenced only 25 times. However, a monotonic rise in the number of citations each year started in 1975 and persisted until 1989, the final year shown on the graph. According to a personal communication from that year by Meier, the needs of applied researchers were "quite well met" by the existing methodology. It wasn't until computers were invented and clinical researchers' mathematical sophistication increased that the Kaplan-Meier method became the standard, but it did grow in importance.
There has been some debate around the Kaplan-Meier approach, despite its widespread use. In his 1983 article "What price Kaplan-Meier?", Miller said

inefficient Kaplan-Meier estimator and advocated for the employment of parametric assumptions by analysts in all cases. A response to the study was given by Meier (2001) in a lecture titled 'The price of Kaplan-Meier.' This was eighteen years later. Because of Meier's original belief that Miller's (1983) findings were wrong, he thought references to it would gradually go away. The high citation count of Miller's article served as inspiration for his presentation. Take a look at Meier et al. (2004) too.

(2) With 18,193 citations (1,342 per year),

Cox, D. R. (1972) Regression models and life tables, *Journal of the Royal Statistical Society, Series B*, 34, pp. 187–220.

This article discusses a subject with broad biological applications: regression analysis of censored failure time data. The hazard function in Cox's (1972) model is semiparametric, which offers several benefits over parametric models when it comes to failure time.

An influential work in statistics, this article was published in Kotz & Johnson (1992b, pp. 519 - 542) and was introduced by R. L. Prentice. Reid (1994) provides some intriguing context for this study by D. R. Cox. Curiously, a crucial idea about the statistical analysis approach allegedly occurred to Professor Cox when he was very sick with the illness and could only recollect it with some difficulty afterwards. Some background on the article was also given by Cox (1986).

(3) With 13,108 citations (256 per year),

Duncan, D. B. (1955) Multiple range and multiple *F*-tests, *Biometrics*, 11, pp. 1 – 42.

In March of 1954, during the Joint Meetings of the Institute of Mathematical Statistics and the Eastern North American Region of the Biometric Society, David Duncan introduced his iconic multiple range test, which compares the means of several populations. Although Duncan first focused on proposing multiple F-tests, his more complicated and labor-intensive versions have never gained the same level of popularity as his more user-friendly multiple range test.

Some historical context for this study was provided by Duncan (1977). In addition, he suggested ditching his multiple range test in favor of Duncan's (1975) procedures.

(4) With 9,504 citations (488 per year),

Marquardt, D. W. (1963) An algorithm for least squares estimation of non-linear parameters, *Journal of the Society for Industrial and Applied Mathematics*, 2, pp. 431 – 441.

The Marquardt algorithm proposed in this paper is used to estimate the parameters in a nonlinear model. See Hahn (1995) and Marquardt (1979) for some interesting background information on this paper.

(5) With 8,720 citations (114 per year),

International journal of basic and applied research
www.pragatipublication.com
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

Litchfield, J. T. & Wilcoxon, F. A. (1949) A simplified method of evaluating dose- effect experiments, *Journal of Pharmacological and Experimental Therapeutics*, 96, pp. 99–113.

A quick visual approach was suggested by the writers to estimate the slope of dosage-percent impact curves and the median effective dose. Wilcoxon was cited by Litchfield (1977) for

significant enthusiasm for working together to create the suggested approach. At Wilcoxon's request, Litchfield and Wilcoxon began discussing the technique the moment he arrived at the laboratory where the latter was working. This was before Litchfield had even visited his boss or contacted HR.

(6) With 8,151 citations (1,590 per year),

Bland, J. M. & Altman, D. G. (1986) Statistical methods for assessing agreement between two clinical measurements, *Lancet*, 1 (8476), pp. 307–310.

The authors outlined basic statistical procedures and visual aids first suggested by Altman and Bland (1983) for comparing two sets of measurements taken with different instruments by means of paired data. (You may see the document online at this URL: http://www.users.york.ac.uk/ mb55/meas/ba.htm.) The authors Bland and Altman describe the paper's significance and how it came to be in their publications from 1992 and 1995.

(7) With 6,788 citations (914 per year),

Felsenstein, J. (1985) Confidence limits on phylogenies: an approach using the bootstrap, *Evolution*, 39, pp. 783–791.

Evolutionary biologists get their theoretical framework from phylogeny, the study of the relationships between all recognized taxonomic groups as seen in the evolutionary tree. According to what I. Hoeschele said in a personal conversation, there is a plethora of data from organism sequencing initiatives like the human genome project that could be utilized to apply the methodologies described here to the work of Nei (1972), the author of our Number 13 publication. Insights on gene functions and solutions to evolutionary mysteries may be gleaned from the findings. Phylogenetics is a dynamic area of research within the framework of adapting methods to genome-scale data and conducting comparative genomic analyses. For more insights on this topic, the reader is urged to see Holmes (2003).

Efron (1979) addressed more fundamental statistical problems in his bootstrap study, which was almost cited (1,889 citations, or 156 per year) but fell short of reaching our list. Felsenstein (1985) investigated a potential use of the bootstrap method. Published in Kotz & Johnson (1992b, pp. 519 - 542) was the groundbreaking book in statistics by Efron (1979), with an introduction by R. J. Beran.

International journal of basic and applied research
www.pragatipublication.com
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

(8) With 6,579 citations (126 per year),

Peto, R., Pike, M. C., Armitage, P., Breslow, N. E., Cox, D. R., Howard, S. V., Mantel, N., McPherson, K., Peto, J. & Smith, K. G. (1977) Design and analysis of randomized clinical trials requiring prolonged observation of each patient. Part II. Analysis and examples,*British Journal of Cancer*, 35, pp. 1–39.

This study has an impressive roster of contributors, including main author Sir Richard Peto as well as co-authors Sir David Cox and Nathan Mantel. To the Leukemia Steering Committee of the United Kingdom Medical Research Council, this report serves as the second instalment in a two-part series. Analytical strategies for comparing patient survival times across randomized clinical trials were the primary emphasis of this paper.

(9) With 6,006 citations (422 per year),

These authors proposed a chi-square test with one degree of freedom for testing the association of disease incidence using 2 ~ 2 contingency tables.

(10) With 5,260 citations (300 per year),

Mantel, N. (1966) Evaluation of survival data and two new rank order statistics arising in its consideration, *Cancer Chemotherapy Reports*, 50, pp. 163–170.

In his discussion of underappreciated works, Garfield (1989) identified Mantel (1966) as an example. 'Actually, slow initial rise characterizes nearly everything,' Mantel said in a 1989 personal communication with Garfield, suggesting that he was philosophical about the matter. He also reasoned that his method's slow recognition by statisticians and epidemiologists could be explained by its publication in a cancer journal.

(11) With 4,306 citations (492 per year),

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm (C/R: pp. 22 – 37), *Journal of the Royal Statistical Society, Series B*, 39, pp. 1–22.

For maximum likelihood estimation using data when certain variables are unobserved, the Expectation Maximization (EM) technique is used. The algorithm's many uses, including those involving censored and shortened data, and the abundance of literature on the topic likely account for the high citation count. On the subject of algebra, McLachlan and Krishnan (1997) wrote a highly acclaimed book. Although Dempster, Laird, and Rubin gave the method the name "EM" in their article, it seems that a small number of researchers had been using it for some time before that. These included McKendrick (1926) and Hartley (1958), who had previously introduced the procedure for calculating maximum likelihood estimates for the general case of count data.

(12) With 3,819 citations (32 per year),

Wilkinson, G. N. (1961) Statistical estimations in enzyme kinetics, *Biochemical Journal*, 80,pp. 324–336.

This article detailed the author's experience with weighted linear and nonlinear regression techniques for solving generic enzyme kinetics issues. To show how nonlinear regression works, we utilized the Michaelis-Menten model, which is popular in enzyme kinetics.

(13) With 3,672 citations (142 per year),

Nei, M. (1972) Genetic distance between populations, *The American Naturalist*, 106, pp. 283 –292.

Nei (1972) proposed a measure of genetic distance based on the identity of genes between populations. The measure can be applied to any pair of organisms.

(14) With 3,511 citations (118 per year),
Visit www.ssc.ca/main/about/history/ dunnett_e.html to read an intriguing essay about Professor Dunnett and the development of his work on repeated comparisons against a control. Furthermore, we are grateful to Professor Dunnett for the information he generously shared with us on this research and its advancements. Fortunately, the multivariate-t distribution, which he developed with Bob Bechhofer and Milton Sobel in their work on ranking and selection (Dunnett & Sobel, 1954, 1955), was the right choice for doing multiple comparisons with a control. Like John Tukey and Henry Scheffe before him, Dunnett (1955) framed the issue in terms of concurrent confidence intervals. It is reasonable to wonder whether the more recent articles that have established the present state-of-the-art were neglected, given the extensive usage of multiple comparison approaches by researchers outside of statistics. In fact, one-stage processes aren't as effective as the step-up and step-down methods proposed by Dunnett and Tamhane (1991, 1992), which are similar to forward selection and backward removal in linear regression. For instance, Dunnett & Tamhane (1991) has 31 citations, even though their approaches are better.

(15) With 3,444 citations (280 per year),

Akaike, H. (1974) A new look at the statistical model identification, *IEEE Transactions on*

International journal of basic and applied research
www.pragatipublication.com
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

Automatic Control, 19, pp. 716 – 723.

Akaike first suggested what is now known as Akaike's Information Criterion (AIC) in this study as a way to estimate a model's dimensionality. The number of citations to this publication is more than triple that of Akaike (1973), a seminal work in statistics that was cited in Kotz & Johnson (1992a, pp. 599 - 624), with a commentary by J. de Leeuw.

(16) With 2,837 citations (376 per year),

Liang, K.-Y. & Zeger, S. (1986) Longitudinal data analysis using generalized linear models, *Biometrika*, 73, pp. 13 – 22.

With an introduction by P. J. Diggle, this work was reissued by Kotz & Johnson (1997, pp. 463 - 482) as a seminal work in statistics. In their 1986 publication, Liang and Zeger addressed longitudinal studies using counts as the response measurement. A widely utilized approach known as generalized estimating equations (GEE) was devised by them.

(17) With 2,810 citations (22 per year),

Cutler, S. J. & Ederer, F. (1958) Maximum utilization of the life table method in analyzingsurvival, *Journal of Chronic Diseases*, 8, pp. 699 – 712.

The writers laid forth the reasoning and computational aspects of the actuarial or life-table approach to analyzing patient survival data. This strategy incorporates all data on survival that has been collected up to the study's end date. According to Cutler (1979), the idea for the article occurred to him around 5 a.m. when he and Ederer were staying in the same hotel room at a scientific symposium. As soon as he had an idea, he roused Ederer to talk about it. It was also said by Cutler (1979) that the work was not a new development in methodology. The writers proved that the life-table technique could be used to retrieve the greatest quantity

derived from the newly structured cancer reporting system's data collection process.

(18) With 2,764 citations (240 per year),

This paper was included by Kotz & Johnson (1997, pp. 123 – 126) as a breakthroughpaper with a discussion by P. J. Huber. Geman & Geman (1984) modified Markovchain Monte Carlo methods and applied them to Bayesian models for the computation of posterior probabilities.

(19) With 2,529 citations (120 per year),

Box, G. E. P. & Cox, D. R. (1964) An analysis of transformations, *Journal of the Royal Statistical Society, Series B*, 26, pp. 211 – 243 (discussion pp. 244 – 252).

From an interview with Professor Box, DeGroot (1987) derived some intriguing background information for this work. For instance, Box said that while serving on a Royal Statistical Society committee, he and Cox were approached by many individuals about the possibility of collaborating. It was partly due to the similarity of their last names that they were inspired to write the article.
The dependent variable in a regression model is usually transformed using a set of power transformations introduced by Box and Cox (1964) in an effort to satisfy the requirements of homoscedasticity and normality of the error terms. The quality of the fit before transforming the dependent variable may be recovered by applying the same transformation to the right side of the model.

(20) With 2,512 citations (76 per year),

Mantel, N. (1963) Chi-square tests with one degree of freedom: extensions of the Mantel–Haenszel procedure, *Journal of the American Statistical Association*, 58, pp. 690 – 700.

The author made two additions to the techniques presented in Mantel & Haenszel (1959), which is number nine on our list: first, the methods may be used to studies that are not retrospective in nature; and second, the number of levels of the research element of interest can be more than two.

(21) With 2,456 citations (46 per year),

Dunnett, C. W. (1964) New tables for multiple comparisons with a control, *Biometrics*, 20, pp.482 – 491.

In this paper, exact critical values are given for the method of Dunnett (1955), Number14 on our list, when two-sided comparisons are made with a control.

(22) With 2,302 citations (42 per year),

Kramer, C. Y. (1956) Extension of multiple range tests to group means with unequal numbersof replications, *Biometrics*, 12, pp. 307 – 310.

Kramer (1956) proposed an approximate method for extending multiple range tests to cases for which the sample sizes are unequal. Kramer’s work was strongly related to the methodology proposed by John Tukey in 1953, whose work was not published. Never- theless, because of the close connection, Kramer's method for the unbalanced case is known as the Tukey– Kramer procedure. (See Benjamini & Braun, 2002, for a discussionof this issue.)

(23) With 2,248 citations (72 per year),

Fisher, R. A. (1953) Dispersion on a sphere, *Proceedings of the Royal Society of London,Series A*, 217, pp. 295–305.

According to Fisher (1953), a significance test called "the analogue of "Student's test" in the Gaussian theory of errors" was constructed from a theory of mistakes that is thought to be suitable for sphere measurements. Here at http://www.library.adelaide.edu.au/digitised/fisher/249.pdf you may see the document in its entirety. This article was Fisher's most referenced work from 1961 to 1975, although having just 277 citations (Garfield, 1977).

(24) With 2,219 citations (240 per year),

Schwarz, G. (1978) Estimating the dimension of a model, *Annals of Statistics*, 6, pp. 461–464.

Schwartz's Bayesian Information Criterion (BIC), introduced in this paper, is a criterionfor model selection that is often mentioned with Akaike's AIC criterion.

(25) With 2,014 citations (382 per year),

Weir, B. S. & Cockerham, C. C. (1984) Estimating *F*-statistics for the analysis of population structure, *Evolution*, 38(6), pp. 1358–1370.

Professor Weir told us that as interest in genetic population structure has grown across many study groups, the number of citations to this work has increased year since its publication. Among these communities you will find forensic scientists, environmentalists, and conservationists.

Comments on the Top 25 List

When compared to the most-cited scientific publications, the most-cited statistics studies do well. Based on the Science Citation Index data from 1945–1988, Garfield (1990) determined the top 100 most-cited publications. Numbers 24, 29, 55, 92, and 94 were assigned, respectively, by Duncan (1972), Litchfield & Wilcoxon (1949), Kaplan & Meier (1958), Marquardt (1963), and Cox (1972). According to Garfield's research, 4,756 citations were given to Kaplan & Meier (1958) while 3,392 citations were given to Cox (1972).

The publications that make up our list were all released before 1987. In Figure 1 we can see a dotplot of the years in which the 25 articles that made up our list were published.

The relationship between a paper's field and its citation count is undeniable. The abundance of

International journal of basic and applied research
www.pragatipublication.com
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

biostatistics-related articles on our list proves this. Just as Kruse (2002) named 27 "highly cited authors in mathematics and statistics" in AmStat News,
Out of all the researchers that had papers published between 1991 and 2001 that were referenced within that time frame, four of them were biostatisticians.
It is reasonable to assume that there is a substantial correlation between the impact of a work and its citation count if we follow the historical trajectory of statistical methods and theory (Efron, 2001, focusing on 1950–1980). Even though they have a lot of citations, just a few of the most seminal works in statistics really made it into our list. According to Efron (2001), the most influential approaches in biostatistics are non-parametric and robust, then the Kaplan-Meier and Cox's methods, and lastly, logistic regression and generalized linear models (GLM). The fact that the two most-cited publications in statistics are Kaplan & Meier (1958) and Cox (1972) has been known for a while. For instance, go to Stigler (1994).

Our list only includes a small selection of the seminal works in the history of statistics. Although they claim to represent "break-through papers in statistics," Kotz and Johnson (1992a, 1992b, 1997) only feature five. The subject of numerous comparisons is covered in four of our most quoted papers: Duncan (1955), Kramer (1956), and Dunnett (1955, 1964). While multiple comparison approaches saw extensive use in statistical practice, they have had little impact on the discipline as a whole. For instance, Tukey (1991) termed the approach of Duncan (1955) a "distraction" and played down its significance. Counting just citations that appeared in statistical journals would be a better way to quantify influence with regard to the discipline of statistics.
Among our 25 publications, four have Nathan Mantel as an author or co-author, which is an intriguing fact. Check out his obituary in July 2002's issue of AmStat News (pp. 35–36) or his website at http://members.aol.com/ savilon/nmantel.html for more details on his accomplishments in the field of statistics. Prior to his retirement in 1974, he devoted a great deal of his career to the National Cancer Institute. He was heavily involved in the consulting industry, which is where many of his groundbreaking scientific contributions began. Three of the publications we cited have Sir D. R. Cox as an author or co-author. To get more information about his background, see Reid (1994).

Most-cited Papers Published Since 1993

We also looked at the most-cited statistical articles released in 1993 and thereafter to get a sense of how the field has evolved over the years. The ten most-cited statisticians, as listed in the November 2003 AmStat News (also see http://www.in-cites.com/top/2003/third03-math.html), were consulted to compile the citation counts for all papers written during this time period. This allowed us to compile a list of the fifteen most-cited papers. With 1,354 citations apiece, the following authors deserve special mention: David L. Donoho, Iain M. Johnstone, Adrian E. Raftery, Adrian F. M. Smith, Peter Hall, Donald B. Rubin, Jianqing Fan, Robert E. Kass, Gareth

International journal of basic and applied research
www.pragatipublication.com
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

O. Roberts, and Siddhartha Chib. The following statistical journals were found to have the most citations in 2002 according to ISI Journal Citation Reports in the Statistics and Probability category (number of citations is in parentheses) for all works published in 1993 or later: Technometrics (2,514), Statistics in Medicine (4,755), Journal of the Royal Statistical Society, Series B (9,458), Annals of Statistics (5,566), Econometrica (9,458), Biometrics (7,469), Biometrika (6,742), and Journal of the American Statistical Association (11,318). (We did not include the 3,626 citations in Fuzzy Sets and Systems in our search.) There may have been some documents that were missed.

The following is our list:

1. Breslow, N. E. & Clayton, D. G. (1993) Approximate inference in generalized linear mixed models, *Journal of the American Statistical Association*, 88, pp. 9 – 25. (558citations)
2. Tierney, L. (1994) Markov-chains for exploring posterior distributions, *Annals of Statistics*, 22, pp. 1701 – 1728. (541 citations)
3. Kass, R. E. & Raftery, A. E. (1995) Bayes Factors, *Journal of the American Statistical Association*, 90, pp. 773 – 795. (533 citations)
4. Donoho, D. L. & Johnstone, I. M. (1994) Ideal spatial adaptation by wavelet shrink- age, *Biometrika*, 81, pp. 425 – 455. (480 citations)
5. Smith, A. F. M. & Roberts, G. O. (1993) Bayesian computation via the Gibbs samplerand related Markov-chain Monte-Carlo methods, *Journal of the Royal Statistical Society, Series B*, 55, pp. 3 – 23. (444 citations)
6. Green, P. J. (1995) Reversible jump Markov-chain Monte Carlo computation and Bayesian model determination, *Biometrika*, 82, pp. 711 – 732. (479 citations)
7. Benjamini, Y. & Hochberg, Y. (1995) Controlling the false discovery rate – a prac- tical and powerful approach to multiple testing, *Journal of the Royal Statistical Society, Series B*, 57, pp. 289 – 300. (294 citations)
8. Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. & Picard, D. (1995) Wavelet shrinkage – asymptopia, *Journal of the Royal Statistical Society, Series B*, 57, pp. 301 – 337. (293 citations)
9. Donoho, D. L. (1995) De-noising by soft thresholding, *IEEE Transactions on Information Theory*, 41, pp. 613 – 627. (292 citations)
10. Grambsch, P. M. & Therneau, T. M. (1994) Proportional hazards tests and diagnosticsbased on weighted residuals, *Biometrika*, 81, pp. 515 – 526. (261 citations)
11. Donoho, D. L. & Johnstone, I. M. (1995) Adapting to unknown smoothness via wavelet shrinkage, *Journal of the American Statistical Association*, 90, pp. 1200 – 1224. (257 citations)

International journal of basic and applied research
www.pragatipublication.com
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

12. Bound, J., Jaeger, D. A. & Baker, R. M. (1995) Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak, *Journal of the American Statistical Association*, 90,pp. 443–450. (252 citations)
13. Albert, J. H. & Chib, S. (1993) Bayesian analysis of binary and polychotomous responsedata, *Journal of the American Statistical Association*, 88, pp. 669–679. (246 citations)
14. Stock, J. H. & Watson, M. W. (1993) A simple estimator of cointegrating vectors in higher-order integrated systems, *Econometrica*, 61, pp. 783–820. (244 citations)
15. Chib, S. & Greenberg, E. (1995) Understanding the Metropolis–Hastings algorithm, *The American Statistician*, 49, pp. 327–335. (240 citations)

The most cited papers presented here tend to be on topics related to Bayesian methods and wavelets, although the topics of multiple testing and proportional hazards modelling are represented. It is interesting to note that it often takes quite a few years for the numberof citations of a paper to reach its maximum rate. A number of the 25 overall most-cited papers are cited now at much higher rates than the most-cited papers of the last decade.

Conclusions

We find the study of citation counts to be very interesting. It is surprising that relativelylittle research has been done on citation counts, rates and patterns in the field of statistics.

Garfield (1979: 16) described early work in this area, including the Citation Index for Stat-istics and Probability, 'a cumulative one-time effort that covers the journal literature of thefield from its inception, early in the twentieth century, through 1966'. This was compiled by John Tukey and published in 1973 as part of the 'Information Access Series' of R&D Press. It provided comprehensive coverage of 40 statistics journals and selective coverageof an additional 100 journals.

In our view it would be very interesting to examine a list of the most-cited papers in each of the top statistics journals (see Campbell & Julious, 1994) or in different applicationareas of statistics. Also, it would be useful to identify papers projected to enter the top 25 most-cited statistical papers and, more generally, 'hot papers' that have attained unusuallyhigh citation rates shortly after publication. For more on this latter topic, the reader is referred to Garfield (2000).