11-14-2012

Appendix A: Statistical Analysis of the Data, Susan Nevelow Mart Study of Search Functions in Lexis and Westlaw

Jeffrey T. Luftig
University of Colorado Engineering Management Program

Jeffrey T. Luftig[1]

Appendix A
Statistical Analysis of the Data

Susan Nevelow Mart Study of Search Functions in Lexis and Westlaw for the article
The Case for Curation: the Relevance of Digest and Citator Results in Westlaw and Lexis

The first step in the analysis of the data was to determine whether the two sets of search functions (Digest Functions Set 1 : KN, LT, and MLTH; Citator Functions Set 2: Shepard's and Keycite) varied statistically in their ability to identify citations that were relevant (to some degree; a degree of relevance analysis followed) to the cases reviewed. For the search engines KN, LT, and MLTH, a total of 1,464, 1,579, and 1,645 citations were identified, respectively, for the cases employed in the study. Each of these citations was then assessed as being 'Relevant' or 'Not Relevant'. The null and alternate hypotheses then tested were:

$$H_0 : \pi_{KN} = \pi_{LT} = \pi_{MLTH}$$
$$H_0 : \pi_{KN} \neq \pi_{LT} \neq \pi_{MLTH}$$

Where π = the proportion of relevant citations identified. To test these hypotheses, a chi-square test of independence (equality for proportions) was conducted. The results of this analysis showed that the null hypothesis should be rejected ($p = 0.000$):

[1] *Lockheed Martin Professor of Management & Program Director*, University of Colorado Engineering Management Program.

Citation Relevancy * Search Function Employed Crosstabulation

			Search Function Employed			Total
			KN	LT	MLTH	
Citation Relevancy	1 Relevant	Count	904_a	579_b	790_c	2273
		% within Search Function Employed	61.7%	36.7%	48.0%	48.5%
		Residual	194.2	-186.6	-7.6	
	2 Not Relevant	Count	560_a	1000_b	855_c	2415
		% within Search Function Employed	38.3%	63.3%	52.0%	51.5%
		Residual	-194.2	186.6	7.6	
Total		Count	1464	1579	1645	4688
		% within Search Function Employed	100.0%	100.0%	100.0%	100.0%

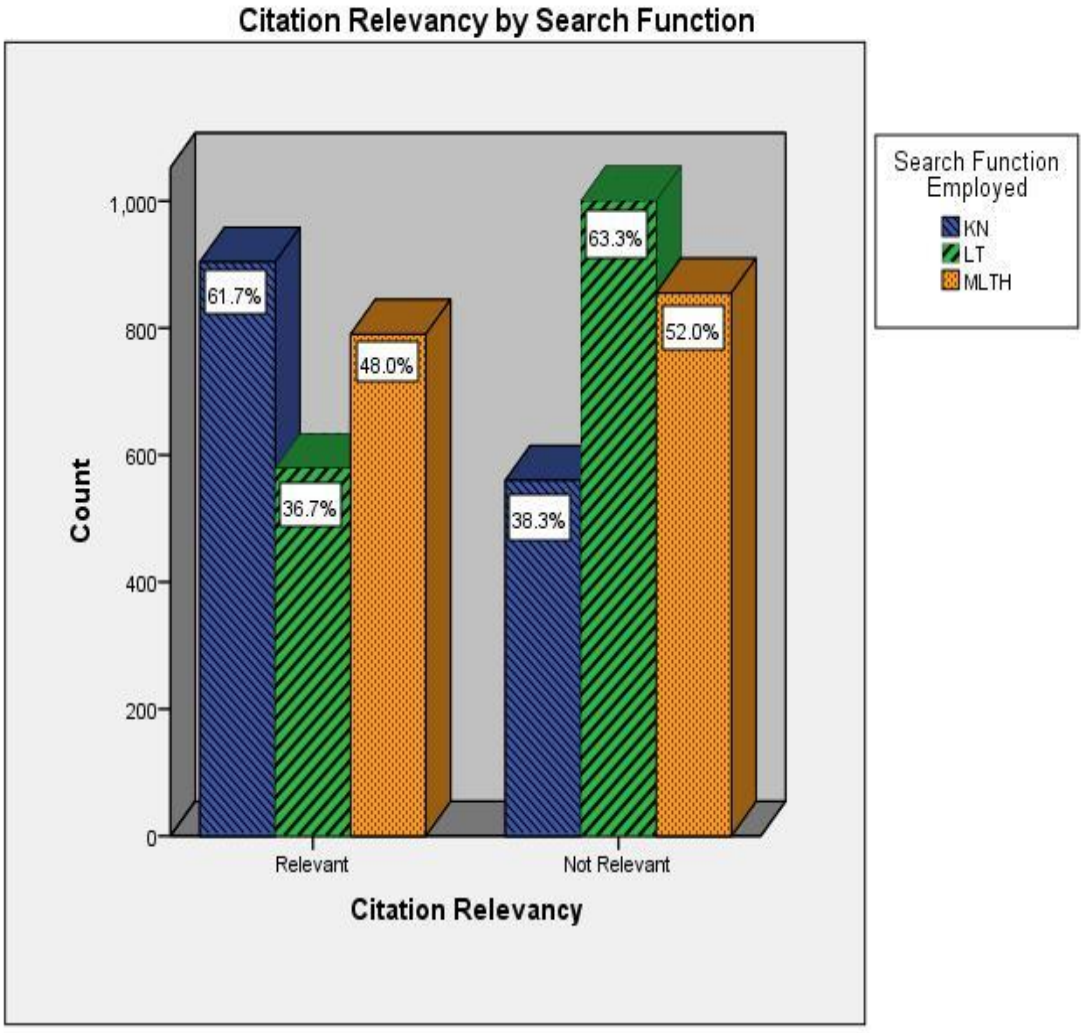
Each subscript letter denotes a subset of Search Function Employed categories whose column proportions do not differ significantly from each other at the .05 level.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	191.522[a]	2	.000
N of Valid Cases	4688		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 709.83.

The differences among the three digest search functions identified may be illustrated with a clustered bar chart:



A widely accepted statistical measure associated with the *importance*, versus significance, of the differences in the proportions observed for a 2 x 3 table is Cramer's V. In this case, the importance of the effect noted was calculated as 0.202, which represents relatively low statistical importance:

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Cramer's V	.202	.000
N of Valid Cases		4688	

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

The results of this analysis indicates that all three search functions were unequal; this does not automatically imply that two of the search engines might still be statistically equal. Utilizing only the results for the two search engines with the closest proportions, and (conveniently) with the smallest sample sizes, a secondary analysis revealed that when considering LT and MLTH alone (excluding the KN data):

$$H_0 : \pi_{LT} = \pi_{MLTH}$$
$$H_0 : \pi_{LT} \neq \pi_{MLTH}$$

the null hypothesis would still be rejected, and the alternate accepted at a significance level (p) equal to 0.000:

Citation Relevancy * Search Function Employed Crosstabulation

			Search Function Employed		
			LT	MLTH	Total
Citation Relevancy	1 Relevant	Count	579$_a$	790$_b$	1369
		% within Search Function Employed	36.7%	48.0%	42.5%
		Residual	-91.5	91.5	
	2 Not Relevant	Count	1000$_a$	855$_b$	1855
		% within Search Function Employed	63.3%	52.0%	57.5%
		Residual	91.5	-91.5	
Total		Count	1579	1645	3224
		% within Search Function Employed	100.0%	100.0%	100.0%

Each subscript letter denotes a subset of Search Function Employed categories whose column proportions do not differ significantly from each other at the .05 level.

Because the data now conformed to a 2 x 2 contingency table analysis, Fisher's Exact Test and *phi* (ϕ) were now employed to test the statistical significance and estimate the importance of the observed differences; respectively:

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	42.522[a]	1	.000		
Continuity Correction[b]	42.058	1	.000		
Fisher's Exact Test				.000	.000
N of Valid Cases	3224				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 670.49.

b. Computed only for a 2x2 table

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	.115	.000
N of Valid Cases		3224	

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

As shown by this portion of the analysis, the differences in the proportion of relevant citations found between LT and MLTH were significant, but of relatively low statistical importance. Summarizing the results of the initial and secondary analyses for the first group of search engines, we would reject the null hypothesis:

$$H_0 : \pi_{KN} = \pi_{LT} = \pi_{MLTH}$$

And infer that: $\pi_{KN} > \pi_{MLTH} > \pi_{MLTH}$ where π represents the proportion of relevant citations identified.

This analysis was then repeated, but for the second set of two search functions– Shepard's and KeyCite. Testing the null and alternate hypotheses:

$$H_0 : \pi_{Shepard's} = \pi_{KeyCite}$$

$$H_0 : \pi_{Shepard's} \neq \pi_{KeyCite}$$

Again, we employed Fisher's Exact Test and *phi* (ϕ) for the analyses of these data. The results appeared as follows:

Citation Relevancy * Search Function Employed Crosstabulation

			Search Function Employed		Total
			Shepard's	KeyCite	
Citation Relevancy	1 Relevant	Count	464$_a$	310$_b$	774
		% within Search Function Employed	43.2%	28.0%	35.5%
		Residual	83.0	-83.0	
	2 Not Relevant	Count	610$_a$	798$_b$	1408
		% within Search Function Employed	56.8%	72.0%	64.5%
		Residual	-83.0	83.0	
Total		Count	1074	1108	2182
		% within Search Function Employed	100.0%	100.0%	100.0%

Each subscript letter denotes a subset of Search Function Employed categories whose column proportions do not differ significantly from each other at the .05 level.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	55.227[a]	1	.000		
Continuity Correction[b]	54.564	1	.000		
Fisher's Exact Test				.000	.000
N of Valid Cases	2182				

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 380.97.

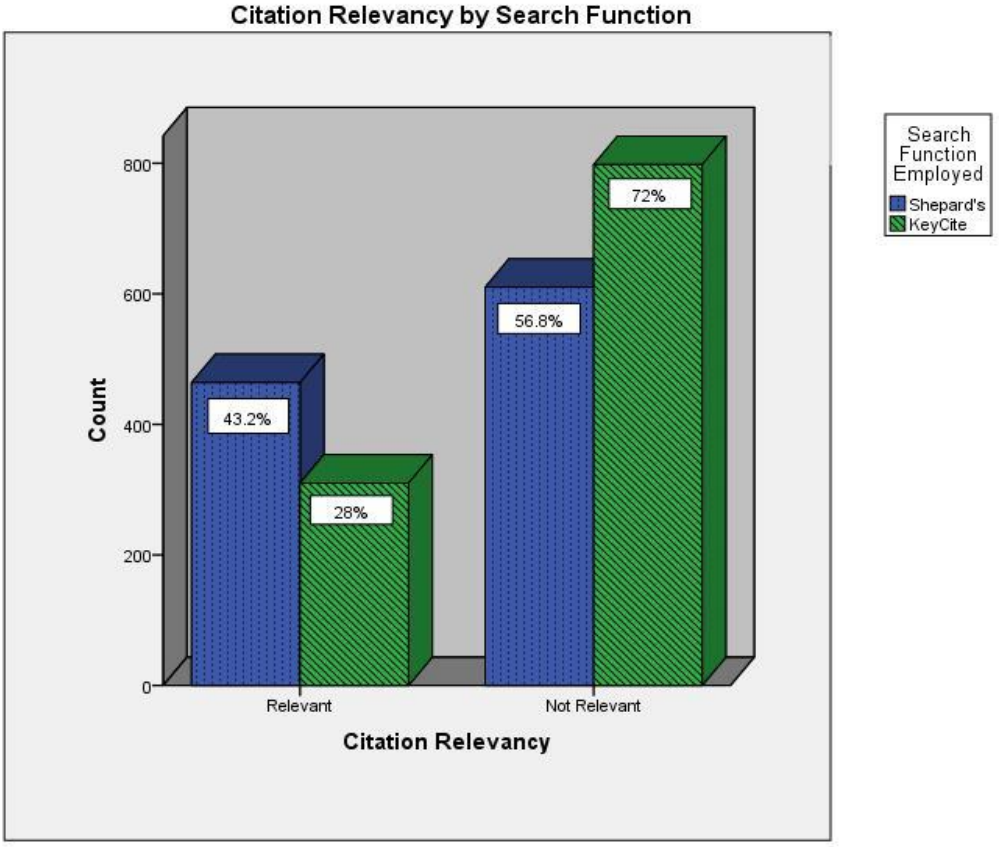
b. Computed only for a 2x2 table

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	.159	.000
N of Valid Cases		2182	

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.



Based on these results, we would reject the null hypothesis (p = 0.000), and infer that:

$$H_0 : \pi_{KeyCite} > \pi_{Shepard's}$$
$$H_0 : \pi_{KeyCite} > \pi_{Shepard's}$$

Although at, again, a relatively low level of statistical importance.

The second step in this analysis was to determine whether, among those citations deemed to be relevant to some extent, the *relative effectiveness* of the identified citations were equivalent. In order to accomplish this analysis, it was first necessary to determine whether the panel of judges (i.e. raters) employed for the analysis of the relative relevance of the citations identified were *concordant*; statistically speaking. In this context, we are referring to a determination of whether the association among k sets of rankings of N objects or specimens shows agreement beyond that which would be expected by chance alone (Siegel, 1956). When these rankings are provided by multiple judges, the common method for describing the concordance among the judges is to employ Kendall's *(W)* Coefficient of Concordance, which is closely related to the average of the Spearman rho (r_s) among the k rank orders (Hays, 1973). When the researcher desires to compare differences among specimens as ranked by multiple judges or raters, the calculation of Kendall's *W* is often thought of as a necessary pre-test, in that a lack of concordance would render tests such as the Friedman ANOVA or Wilcoxon Signed-Rank tests on the n specimens questionable.

The null and alternative hypotheses tested for the Kendall test of concordance (Sheskin, 1997) may be stated as:

$$H_0 : W = 0$$
$$H_1 : W \neq 0$$

In this case, acceptance of the null hypothesis would imply that there was no concordance among the k judges or raters for the N specimens (in this case, citations) evaluated.

The five search functions were tested for this study, and each of the five (5) judges (SPSSPc reports each judge or rater as an 'N' in the summary tables which follow) assessed five (5) randomly selected relevant citations generated for each of five (5) cases. As a result, each of the five judges evaluated the same 75 citations generated by the search functions. The result of Kendall's concordance analysis was statistically significant (p = 0.000):

Test Statistics

N	5
Kendall's W[a]	.590
Chi-Square	218.295
df	74
Asymp. Sig.	.000

a. Kendall's Coefficient of Concordance

The null hypothesis was rejected, leading the researcher to find that sufficient statistical evidence exists to infer that the five judges were concordant (in agreement) in their evaluations of the degree of relevancy exhibited by the citations generated by the five search functions.
An interesting feature of this study was that the degree of relevance for the citations identified by the five search functions were not to be compared in total; that is, the results of the three digest search functions - MLTH, KN, and LT - were to be compared; followed by a mutually exclusive comparison of the results for the two citator functions -Shepard's versus Keycite. It would be unlikely but possible that the concordance exhibited among the judges could correspond to the one, but not both, of the search engine results. To confirm that this was not the case, Kendall's test of concordance was executed for the 45 citations identified by the LT, KN, and MLTH search engines:

Test Statistics

N	5
Kendall's W[a]	.571
Chi-Square	125.618
df	44
Asymp. Sig.	.000

a. Kendall's Coefficient of Concordance

and then repeated for the Shepard's and Keycite citations:

Test Statistics

N	5
Kendall's W[a]	.581
Chi-Square	84.204
df	29
Asymp. Sig.	.000

a. Kendall's Coefficient of Concordance

As shown by these results, we would reject the null hypothesis that no concordance existed among the five (5) judges whether the data were evaluated for all citations identified by the five search functions, or within the two different sets of search functions to be compared.

Having established concordance among the judges, the final step in assessing the relative *degree of relevance* among the citations identified by the two sets of search engines was executed. In order to compare the KN, LT, and MLTH search functions, and given that a univariate (i.e. single dimensional) ordinal scale was employed by the judges to assess the relative relevancy of

the citations generated, the median value for the five (5) ratings generated by the judges was generated for each citation. Given that these data represented dependent data, the Friedman Analysis of Variance by Ranks was selected as the most robust test for this comparison. The null and alternative hypotheses for this test may be stated as suggested by Hays, 1973:

H_0 : The degree of relevance for the relevant citations identified by the three search engines are equivalent

H_1 : The degree of relevance for the relevant citations identified by the three search engines are not equivalent

Although some statisticians (Sheskin, 1997) would express the hypotheses using the population medians (θ):

$$H_0 : \theta_{KN} = \theta_{LT} = \theta_{MLTH}$$
$$H_1 : \theta_{KN} \neq \theta_{LT} \neq \theta_{MLTH}$$

Conducting the Friedman ANOVA for the three search engines, we find that we would accept the null hypothesis ($p = 0.687$):

Descriptive Statistics

	N	Average Median	Minimum	Maximum
KN	10	3.70	2	5
LT	10	3.90	1	5
MLTH	10	3.60	1	5

Ranks

	Mean Rank
KN	1.90
LT	2.20
MLTH	1.90

Test Statistics[a]

N	10
Chi-Square	.750
df	2
Asymp. Sig.	.687

a. Friedman Test

The result of this analysis would lead us to infer that the relevant results returned by the KN, LT, and MLTH search functions were equally useful in terms of the relevancy of the citations identified.

Moving on to the analysis of the efficacy of the Shepard's and Keycite search functions, we would again test the hypotheses that:

H_0 : The degree of relevance for the relevant citations identified by the two search engines are equivalent
H_1 : The degree of relevance for the relevant citations identified by the two search engines are not equivalent, or:

$$H_0 : \theta_{\text{Shepard's}} = \theta_{\text{Keycite}}$$
$$H_1 : \theta_{\text{Shephard's}} \neq \theta_{\text{Keycite}}$$

In this application, as we have only two groups, the appropriate test would be the Wilcoxon Signed-Rank Test. Comparing the two sets of medians:

Ranks

		N	Mean Rank	Sum of Ranks
KeyCite - Shepard's	Negative Ranks	7[a]	4.64	32.50
	Positive Ranks	1[b]	3.50	3.50
	Ties	2[c]		
	Total	10		

a. KeyCite < Shepard's

b. KeyCite > Shepard's

c. KeyCite = Shepard's

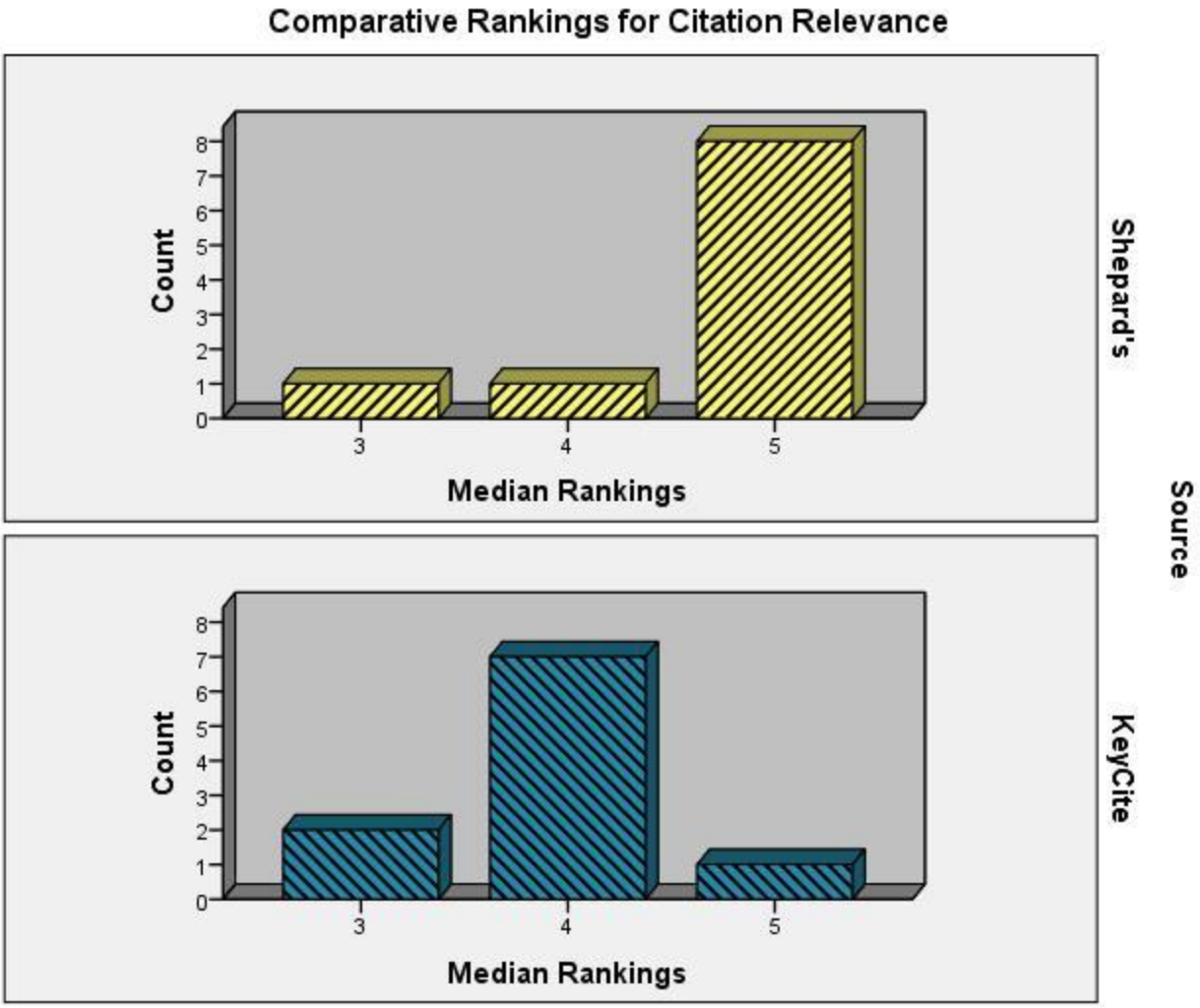
Test Statistics[a]

	KeyCite - Shepard's
Z	-2.126[b]
Asymp. Sig. (2-tailed)	.033

a. Wilcoxon Signed Ranks Test

b. Based on positive ranks.

We would reject the null hypothesis ($p = 0.33$) at an $\alpha = 0.05$, and infer that there was a statistically significant difference in the relative relevancy of the relevant citations identified by these two search engines. Illustration 1 which follows reflects the differences observed:



The researcher found the relevant citations identified by Shepard's to be more relevant to the cases employed than the relevant citations identified for those same cases by Keycite, based on the median values associated with each citation, generated by the concordant judges. It should be noted, however, that the fact that the values are statistically significantly different does not automatically imply that the difference observed is *important*. Using a square root transformation of the data as suggested by Dixon and Massey (1983), and conducting a one way ANOVA to generate the required data, the approximate omega-squared (ω^2) value (Hays, 1973) for the difference between the two groups was 24.6%; representing a relatively low level of importance.

References:

Dixon, W. J. & Massey, F. J. *Introduction to Statistical Analysis*; 4th Edition, McGraw-Hill Book Co., New York, 1983

Hays, W. L. *Statistics for the Social Sciences*; Holt, Rinehart, and Winston, New York, 1973

Sheskin, D.J. *Handbook of Parametric and Nonparametric Statistical Procedures*; CRC Press, New York, 1997

Siegel, S. *Nonparametric Statistics for the Behavioral Sciences*; McGraw-Hill Book Co., New York, 1956