The Magical Influene of Statistical Significance

Abdalla Alsmadi* **Ferial Al-Hajj****

Abstract

This paper examined 1122 statistical tests found in 55 master theses accredited during 1995-2000 at Mu'tah University. It tried to answer two main questions: First, do researchers still relying on the level of significance (α) as the only criterion to judge the importance of relations and differences? Second, to what extent practical significance can be found along with statistical significance? Results showed that researchers do consider statistical significance as the only criterion to judge the importance of their findings. 74.33% of the statistically significant tests were having a small practical significance, and only 10.27% were of large practical significance.

* Dept. of Psychology, Mu'tah University.
** Ministry of Education .

The Magical Influene of Statistical Significance

Abdalla Alsmadi* **Ferial Al-Hajj****

This paper examined 1122 statistical tests found in 55 master theses accredited during 1995-2000 at Mu'tah University. It tried to answer two main questions: First, do researchers still relying on the level of significance (α) as the only criterion to judge the importance of relations and differences? Second, to what extent practical significance can be found along with statistical significance? Results showed that researchers do consider statistical significance as the only criterion to judge the importance of their findings. 74.33% of the statistically significant tests were having a small practical significance, and only 10.27% were of large practical significance.

Tests of significance are widely used in social sciences research. Yet there is a considerable difference of opinion on their value in achieving the goals of science. Questioning the utility of tests of significance is not new. There is a long and honorable tradition of blistering attacks on the role of significance testing in the behavioral sciences.

Testing a particular belief or assumption (which we can call a hypothesis) follows a pattern. We assume something (the hypothesis) to be true, and we test this hypothesis by comparing observations (data) on the real world with what our hypothesis would lead us to

* **Dept. of Psychology, Mu'tah University.**
** **Ministry of Education .**

expect. If we find the real world corresponds closely enough with what our hypothesis led us to expect, we continue to believe our hypothesis .If what we observe dose not correspond closely enough to what we expect, and we suspect our hypothesis of being false. (Ramon, 1976)

A debate has continued for more than forty years about significance test that can legitimately be performed. One tradition declares that when the data are measured on an ordinal scale, parametric statistics should not be used. The other maintains that only mathematical characteristics of the numbers have to be satisfied in order to justify parametric methods. Two errors in particular have made consensus difficult to reach. Firstly, some writers fail to distinguish between statistical treatment of the data, considered merely as numbers, and the use of numerical results to justify statements about the world. Secondly, statistical significance is sometimes treated as a property of the world or of the date whereas in fact it is a numerical answer to one question out of many alternative questions that might be asked about the data. Each question has a different, correct answer. These errors often lead to misinterpretation of what classical significance tests actually tell us about the world. (MacRae, 1988)

Three habits of language usage are usually make the unconscious misinterpretations. Overcoming theses habits will help avoid that problem. First, don't stop at saying "significant" but you need to say "statistically significant". Second, don't say things like "my results approached statistical significance". Finally, don't say things like "the statistical significance testing evaluated whether the results were due to chance" simply because this language gives the impression that replicability is evaluated by significance testing, which is not true. (Thompson, 2003). This seems to contradict Ramon's opinion in which he says that significance testing allows us

to evaluate differences between what we expect on the basis of our hypothesis, and what we observe, but only in terms of one criterion, the probability that these differences could have occurred by "chance".

Although significance testing provides criteria for determining when expectations and observations are in agreement or disagreement by considering the effect of chance factors in a rigorous fashion, it dose not completely eliminate the subjective aspects of making this assessment . Subjective judgments are still a part of the process, since the choice of the probability which will lead to a decision that the hypothesis is false is an arbitrary, subjective choice. Thus testes of significance introduce a certain amount of rigor and eliminate a certain amount of subjectivity, but do not remove all subjectivity from the process. (Ramon, 1976)

To make it clearer, suppose that a researcher administers a continuous individual difference measure, say an intelligence test and then promptly proceeds to perform a median split on the measure, divides the sample into two groups on some dependent measures through a *t*-test or ANOVA. This procedure is certainly logical permissible, but it is just as certainly stupid. Not only it results in an avoidable loss of power, but it also regards a subject with an IQ of 99, for example, as being qualitatively different from a subject with an IQ of 101. Analogously, using significance testing to decide whether data are similar to predictions could lead one to conclude incorrectly that a result that was not significant' at p =0.053 was qualitatively different from a result yielding p= 0.049. We would fail miserably in our goal of explaining behavior if we treated all statistically significant results as equal. One may try arguing that all significant results imply similarity between theory and data. We would also fail in our goal of explaining behavior if we regarded a p of 0.051 as being qualitatively different and substantially worse than a p of 0.049. The

binary judgment afforded by significance tests does not yield an optimal solution to the question of similarity because similarity is a continuous variable, not a dichotomous variable. The heart of theory corroboration is determining the degree of fit between one theoretical implications and the obtained data (Harris, 1991)

Schneider and Darcy (cited in Richard, 1988) list seven factors that determine the outcome of significance tests:

1) Actual strength of impact.
2) Number of cases used in the study.
3) Variation among cases on relevant variables
4) The complexity of the analysis (degree of freedom).
5) The appropriateness of the statistical measures and tests used.
6) The hypothesis tested.
7) The significance level chosen.

Note that, only one factor deals with the impact of the outcome. If gender differences were investigated and found significant, it is usually the case to say that there is a significant difference in the dependent variable due to gender, ignoring all other six factors just mentioned.

Among the seven factors affecting the statistical significance listed be Schneider and Darcy, Thompson (1988) notes that the one having the greatest impact is the size of the sample. When working with large samples, virtually all null hypotheses will be rejected, since – as Thompson said - the "null hypotheses of no differences are almost never exactly true in the population" (Palomares, 1987, p.14).

The suggestion though, authors should identify the smallest sample size at which the result would have remained statistically significant (Richard, 1988). Another suggestion is to examine the practical significance by computing the "effect size." index. It can be characterized as the degree to which the phenomenon exists. It is used

by the researcher to garner some insight regarding result importance. The basic question to be answered when conducting research is: "how much of the dependent variable is accounted for by the independent variable?" or "what proportion of the variance in the dependent variable is explained by the observed effect?" (Richard, 1988)

Effect Size Indices

The Effect Size of T–test: (d)

Case 1: Independent samples t-test where $n_1=n_2$ and $\sigma_1^2 = \sigma_2^2$ (H_0: $\mu_1 - \mu_2 = 0$). The effect size (d) is computed by:

$$d = \left| \frac{\overline{X}_1 - \overline{X}_2}{\sigma} \right| \qquad (1)$$

If $\sigma_1^2 \neq \sigma_2^2$ then, σ is substituted by the common variance (σ_{com})

$$\sigma_{com} = \sqrt{\frac{\sigma_1^2 + \sigma_2^2}{2}} \qquad (2)$$

Case 2: dependent samples t-test where (H_0:X_a-X_b=0). The effect size (d) is computed by:

$$d = \frac{\overline{X}_{a-b}}{\sigma_{a-b}} \qquad (3)$$

Where: $\overline{X}_{a-b}$ is the mean of the differences between both samples. σ_{a-b} is the standard deviation of the differences between both samples, and computed by

$$\sigma_{a-b} = \sqrt{\sigma_a^2 + \sigma_b^2 - 2r\sigma_a\sigma_b} \qquad (4)$$

According to Cohen's standards, effect size is considered small when d is 0.0 to 0.49, medium when d is 0.5 to 0.79, and large when d is 0.8 to 1.0, (Cohen, 1977)

Effect size of Multiple Regression (F^2)
The effect size is computed by:

$$F^2 = \frac{R^2}{1-R^2} \qquad (5)$$

Where: R^2 the proportion of variance in the dependent variable explained by the independent variables. According to Cohen's standards, effect size is considered small when $0 \leq F^2 \leq 0.14$,medium when $0.15 \leq F^2 \leq 0.34$,and large when $0.35 \leq F^2$ (Cohen, 1979)

Effect size of Chi Square (W)
Case1: Goodness of fit. The effect size is computed by:

$$w = \sqrt{\sum_{i=1}^{m} \frac{\left(P_{\exp ected} - P_{obtained}\right)^2}{P_{\exp ected}}} \qquad (6)$$

Note that Proportions are used in this equation but not the row values. W=0.0 when the expected and observed proportions are equal and the null is true.

Case 2: Two variables independency. The effect size is computed by:

$$w = phi\,(\phi) = \sqrt{\frac{\chi^2}{N}} \qquad (7)$$

According to Cohen's standards, effect size is considered small $w \leq 0.29$, medium when $0.3 \leq w \leq 0.49$, and large when $w \geq 0.5$ (Cohen, 1979)

The Effect Size of MANOVA (D^2)

The effect size is computed by: $D^2 = \frac{NT^2}{n_1 n_2}$ (8)

Where T^2 is computed by:

$$T^2 = \frac{(N-2)PF}{N-P-1} \qquad (9)$$

P: is number of dependent variables.

According to Steven's standards, effect size is considered small $D^2 \leq 0.49$, medium when $.5 \leq D^2 \leq 0.95$, and large when $D^2 \geq 1.0$ (Stevens, 1980)

Purpose

The purpose of this study is examining the extent to which researchers rely on statistical significance as a criterion to judge the importance of the findings. Furthermore, it aimed at investigating the percent of tests that were statistically significant having at the same time large practical significance.

Methodology

Population

The population of this study consisted of 1122 statistical tests employed by 55 quantitative master thesis credited at the Faculty of Education at Mu'tah University during 1995-2000; (quantitative thesis means any thesis used inferential statistics for hypothesis testing).

Procedures

All thesis were classified into two main categories: quantitative or qualitative. Out of the sixty two thesis credited 1995-2000, only seven (11.3%) were classified as qualitative, and fifty five (88.7%) were classified as quantitative.

Out of the 1122 statistical tests reviewed, 10 t-tests, and 53 tests of ANOVA were excluded because of the incomplete data

required for practical significant calculations. Therefore, this study examined 1059 statistical tests as presented by table 1.

Table (1)

Types and frequencies of statistical tests

Statistical Test	Frequency	Percentage
T-Test	193	18.22%
ANOVA	712	67.23%
Chi – Square	56	5.28%
Multiple regression	82	7.74%
MANOVA	16	1.5%
Total	1059	

The practical significance then was investigated by computing the effect size index for each statistical test as presented by table 2. Then, the effect size value for MANOVA was classified according to Steven's standards. The other effect size values were classified according to Cohen's standards.

Table (2)

Effect size index used for each statistical test

Statistical test	Effect size index
T-test	D
ANOVA	ή
Chi – Square	W
Multiple regression	F^2
MANOVA	D^2

Results

This study tried to answer two main questions: First, do researchers still relying on the level of significance (α) as the only

criterion to judge the importance of relations and differences? Second, to what extent practical significance can be found along with statistical significance?

Results of the first question

A careful review of 55 quantitative thesis revealed that researchers still relying heavily on significance level ($\alpha \leq 0.05$) as the only criterion to judge the importance of relations and differences. Researchers do not use any other criterion to judge the importance of the study. As presented by table (3), the significance level ($\alpha \leq 0.05$) is most used compared with the significance level ($\alpha \leq 0.01$). The only exception is the Multiple Regression procedure in which the opposite is found.

Table (3)

statistical tests according the level of significance used

Statistical Test	Frequency	(α)	Percentage
T-Test	114	0.05	59.07%
	79	0.01	40.93%
ANOVA	515	0.05	72.33%
	197	0.01	27.67%
Chi – Square	50	0.05	89.29%
	6	0.01	10.71%
Multiple regression	24	0.05	29.27%
	58	0.01	70.73%
MANOVA	15	0.05	93.75%
	1	0.01	6.25%
Total	1059		

Results of the second question

To answer the second question, the practical significance was computed for all statistical tests. Then, tests were classified twice

according the statistical significance and the practical significance. Results are presented by tables (4 through 8), in which the first column contains three different value intervals (suggested by Choen , 1977) as criterions for Practical Significance classification.

Table (4)

T - tests classified according statistical and practical significance

Statistical Significance	Significance		Non - Significance		Total	
	Frequency	Percent	Frequency	Percent	Frequency	Percent
Practical Significance						
Small (0.0 - 0.49)	75	60.98%	68	97.14%	143	74.2%
Medium (0.5 – 0.79)	18	14.63%	0.0	0.0	18	9.3%
Large (0.8 – 1)	30	24.39%	2	2.86%	32	16.5%
Total	123	100%	70	100%	193	100%
Total Percent		63.7%		36.3%		

It is clearly noted in table (4) that 60.98% of the statistically significant t-tests were of small Practical significant, and only 24.39% were of large practical significance.

Table (5)

ANOVA tests classified according statistical and practical significance

Statistical Significance	Significance		Non - Significance		Total	
	Frequency	Percent	Frequency	Percent	Frequency	Percent
Practical Significance						
Small (0.0 – 0.24)	247	95.36%	449	99.12%	696	97.8%
Medium (0.25 – 0.39)	6	2.32%	0.0	0.0%	6	0.8%
Large (0.4 – 1)	6	2.32%	4	0.88%	10	1.4%
Total	259	100%	453	100%	712	100%
Total Percent		36.3%		63.7%		

Table (5) shows that 95.36% of the statistically significant ANOVA tests were of small Practical significant, and only 2.32% were of large practical significance

Table (6)

Chi Square tests classified according statistical and practical significance

Statistical Significance	Significance		Non - Significance		Total	
	Frequency	Percent	Frequency	Percent	Frequency	Percent
Practical Significance						
Small (0.0 – 0.29)	13	30.23%	11	84.62%	24	42.8%
Medium (0.3 – 0.49)	9	20.93%	2	15.38%	11	19.7%
Large (0.5 – 1)	21	48.84%	0.0	0.0	21	37.5%
Total	43	100%	13	100%	56	100%
Total Percent		76.8%		23.2%		

The analysis of Chi Square tests revealed that 30.23% of the statistically significant tests were of small Practical significant, and 48.84% were of large practical significance.

Table (7)

Multiple Regression tests classified according statistical and practical significance

Statistical Significance	Significance		Non - Significance		Total	
	Frequency	Percent	Frequency	Percent	Frequency	Percent
Practical Significance						
Small (0.0 – 0.14)	26	43.34%	21	95.45%	47	57.3%
Medium (0.15 – 0.34)	17	28.33%	0.0	0.0%	17	20.7%
Large (0.35 – 1)	17	28.33%	1	4.55%	18	22%
Total	60	100%	22	100%	82	100%
Total Percent		73.2%		26.8%		

Results in table 6 showed that 43.34% of the statistically significant multiple regression tests were of small Practical significant, and 28.33% were of large practical significance.

Table (8)

MANOVA tests classified according statistical and practical significance.

Statistical Significance	Significance		Non - Significance		Total	
	Frequency	Percent	Frequency	Percent	Frequency	Percent
Practical Significance						
Small (0.0 – 0.49)	4	66.66%	9	90%	13	81.1%
Medium (0.5 – 0.95)	1	16.67%	1	10%	2	12.6%
Large (0.96 or above)	1	16.67%	0.0	0.0	1	6.3%
Total	6	100%	10	100%	16	100%
Total Percent		37.6%		62.4%		

It is easily noted that 60.98% of the statistically significant MANOVA tests were of small Practical significant, and only 24.39% were of large practical significance.

Table (9)

A summary for all statistical tests classified according statistical and practical significance.

Statistical Significance	Significance		Non - Significance		Total	
	Frequency	Percent	Frequency	Percent	Frequency	Percent
Practical Significance						
Small	365	74.33%	558	98.22%	923	87.15%
Medium	51	10.38%	3	0.5%	54	5.09%
Large	75	10.27%	7	1.2%	82	7.43%
Total	491	100%	568	100%	1059	100%
Total Percent		46.3%		53.7%		

Discussion

Regarding the first question the study tried to answer, it has been found that non of the 55 thesis reviewed computed any practical significance index along with the statistical significance. All researchers judged the importance of their findings and announced their recommendations according the statistical significance criterion. The answer of the second question was presented in tables 4-8. Results showed that the percent of statistically significant tests with small practical significance was always more than the percent of statistically significant tests with large practical significance except for the chi square test. Table 9, however, gives the general view. Generally, it has been found that 74.33% of the statistically significant tests were having no or small practical significance. large practical significance was found for only 10.27% of tests that were statistically significant. These results may be interpreted as a function of researcher's desirability of seeking statistically significant results by having large sample size.

The question that needs to be answered is why more researchers don't utilize effect size statistics when reporting results? Is it a "*magical influence*" of statistical significance? One reason might be the influence of the significance testing which pushes the attention away from effect size (Moore, 1991). A second reason may be that researchers fail to recognize that all analyses in fact are testing associations, and therefore that effect sizes analogous to the coefficient of determination are appropriate for all analytic techniques. Researchers may be afraid that research design may be viewed as being correlational. A third reason might be a part of the answer is the lack of knowledge about effect size analysis. Master program includes one course in statistics and effect size is not always among the subjects discussed.

Conclusion

Increasingly, researchers note that they obtain statistically significant results, but careful scrutiny of the data demonstrates that such differences are not necessarily meaningful. The use of statistical significance testing in educational research has often misused, they are useful for assuring us that our results are not due to random sampling fluctuation and we should not stop there. Researchers need to recognize the limitations of the significance testing and be wary of studies that base decisions of importance on such testing. Effect size statistics aid in the interpretation of results and will provide a guide to the relative importance of the study.

REFERENCES

- Cohen, J. (1977). Statistical Power Analysis for the Behavioral Sciences (revised edition). Subsidiary of Harcourt Brace Jovanovich Publishers, New York San Francisco, London.
- Harris. Monica J. (1991). Significance Tests Are Not Enough. London, Theory & Psychology © Sage. Vol 1(3):375-382.
- Henkel. R.E. (1976). Tests of Significance. Beverly Hills, CA: Sage Publications.
- MacRae, (1988). Measurement Scales and Statistics: What Can Significance Tests Tell Us About The World? British Journal of Psychology, 79, 161-171.
- Moore, Mary Ann. (1991). The Place of Significance Testing in Contemporary Social Science. Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, April 3-7, 1991). ED 333036.

- Palomares, Ronald S. (1990, Nov). Alternatives to Statistical Significance Testing. Paper presented at the Annual Meeting of the Mid-South Educational Research Association (19th, New Orleans, LA, November 14-16).
- Richard D. Keaster. (1988). Statistical Significance Testing: From Routine to Ritual. Paper presented at the Annual Meeting of the Mid-South Educational Research Association. Louisville, KY, November 9th, 1988.
- Stevens,J.P. (1980). Power of Multivariate Analysis of Variance Test. Psychological Bulletin, 88, 728-737.
- Thompson, Bruce. (2003). The Concept of Statistical Significance Testing.ERIC/AEDigest. http://ericae.net/edo/ed466654.htm, 22.03.2003.

تاريخ ورود البحـــث : ١٤/٤/٢٠٠٣م
تاريخ ورود التعديلات : ٦ /١/٢٠٠٤م
تاريخ القـبول للنشـر : ١٢/٢/٢٠٠٤م