P – VALUE, A TRUE TEST OF STATISTICAL SIGNIFICANCE? A CAUTIONARY NOTE

Tukur Dahiru (MBBS), FMCPH, Dip. HSM (Israel)
Dept. of Community Medicine, Ahmadu Bello University, Zaria, Nigeria.

All Correspondence to:
Dr. Tukur Dahiru
MBBS, FMCPH, Dip HSM (Israel)
DEPT OF COMMUNITY MEDICINE
AHMADU BELLO UNIVERSITY,
ZARIA, NIGERIA.
Email:tukurdahiru@yahoo.com

ABSTRACT

While it's not the intention of the founders of significance testing and hypothesis testing to have the two ideas intertwined as if they are complementary, the inconvenient marriage of the two practices into one coherent, convenient, incontrovertible and misinterpreted practice has dotted our standard statistics textbooks and medical journals. This paper examine factors contributing to this practice, traced the historical evolution of the Fisherian and Neyman-Pearsonian schools of hypothesis testing, exposed the fallacies and the uncommon ground and common grounds approach to the problem. Finally, it offers recommendations on what is to be done to remedy the situation.

INTRODUCTION

The medical journals are replete with P values and tests of hypotheses. It is a common practice among medical researchers to quote whether the test of hypothesis they carried out is significant or non-significant and many researchers get very excited when they discover a "statistically significant" finding without really understanding what it means. Additionally, while medical journals are florid of statement such as: "statistical significant", "unlikely due to chance", "not significant," "due to chance", or notations such as, "P > 0.05", "P < 0.05", the decision on whether to decide a test of hypothesis is significant or not based on P value has generated an intense debate among statisticians. It began among founders of statistical inference more than 60 years ago[1-3]. One contributing factor for this is that the medical literature shows a strong tendency to accentuate the positive findings; many researchers would like to report positive findings based on previously reported researches as "non-significant results should not take up" journal space[4-7].

The idea of significance testing was introduced by R.A. Fisher, but over the past six decades its utility, understanding and interpretation has been misunderstood and generated so much scholarly writings to remedy the situation[3]. Alongside the statistical test of hypothesis is the P value, which similarly, its meaning and interpretation has been misused. To delve well into the subject matter, a short history of the evolution of statistical test of hypothesis is warranted to clear some misunderstanding.

A Brief History of P Value and Significance Testing

Significance testing evolved from the idea and practice of the eminent statistician, R.A. Fisher in the 1930s. His idea is simple: suppose we found an association between poverty level and malnutrition among children under the age of five years. This is a finding, but could it be a chance finding? Or perhaps we want to evaluate whether a new nutrition therapy improves nutritional status of malnourished children. We study a group of malnourished children treated with the new therapy and a comparable group treated with old nutritional therapy and find in the new therapy group an improvement of nutritional status by 2 units over the old therapy group. This finding will obviously, be welcomed but it is also possible that this finding is purely due to chance. Thus, Fisher saw P value as an index measuring the strength of evidence against the null hypothesis (in our examples, the hypothesis that there is no association between poverty level and malnutrition or the new therapy does not improve nutritional status). To quantify the strength of evidence against null hypothesis "he advocated P < 0.05 (5% significance) as a standard level for concluding that there is evidence against the hypothesis tested, though not as an absolute rule" [8]. Fisher did not stop there but graded the strength of evidence against null hypothesis. He proposed "if P is between 0.1 and 0.9 there is certainly no reason to suspect the hypothesis tested. If it's below 0.02 it is strongly indicated that the hypothesis fails to account for the whole of the facts. We shall not often be astray if we draw a conventional line at 0.05" [9]. Since Fisher made this statement over 60 years ago, 0.05 cut-off point has been used by medical researchers worldwide and has become ritualistic to use 0.05 cut-off mark as if other cut-off points cannot be used. Through the 1960s it was a standard practice in many fields to report P values with the star attached to indicate P < 0.05 and two stars to indicate P < 0.01. Occasionally three stars were used to indicate P < 0.001. While Fisher developed this practice of quantifying the strength of evidence against null hypothesis some eminent statisticians where not accustomed to the subjective interpretation inherent in the method [7]. This led Jerzy Neyman and Egon Pearson to propose a new approach which they called "Hypothesis tests". They argued that there were two

types of error that could be made in interpreting the results of an experiment as shown in Table 1.

Result of experiment	The truth	
	Null hypothesis true	Null hypothesis false
Reject null hypothesis	Type I error rate(α)	Power = 1-β
Accept null hypothesis	Correct decision	Type II error rate (β)

Table 1.Errors associated with results of experiment.

The outcome of the hypothesis test is one of two: to reject one hypothesis and to accept the other. Adopting this practice exposes one to two types of errors: reject null hypothesis when it should be accepted (i.e., the two therapies differ when they are actually the same, also known as a false-positive result, a type I error or an alpha error) or accept null hypothesis when it should have rejected (i.e. concluding that they are the same when in fact they differ, also known as a false-negative result, type II error or a beta error).

What does P value Mean?

The P value is defined as the probability under the assumption of no effect or no difference (null hypothesis), of obtaining a result equal to or more extreme than what was actually observed. The P stands for probability and measures how likely it is that any observed difference between groups is due to chance. Being a probability, P can take any value between 0 and 1. Values close to 0 indicate that the observed difference is unlikely to be due to chance, whereas a P value close to 1 suggests no difference between the groups other than due to chance. Thus, it is common in medical journals to see adjectives such as "highly significant" or "very significant" after quoting the P value depending on how close to zero the value is.

Before the advent of computers and statistical software, researchers depended on tabulated values of P to make decisions. This practice is now obsolete and the use of exact P value is much preferred. Statistical software can give the exact P value and allows appreciation of the range of values that P can take up between 0 and 1. Briefly, for example, weights of 18 subjects were taken from a community to determine if their body weight is ideal (i.e. 100kg). Using student's t test, t turned out to be 3.76 at 17 degree of freedom. Comparing t_{stat} with the tabulated values, t= 3.26 is more than the critical value of 2.11 at p=0.05 and therefore falls in the rejection zone. Thus we reject null hypothesis that ì = 100 and conclude that the difference is significant. But using an SPSS (a statistical software), the following information came when the data were entered, t = 3.758, P = 0.0016, mean difference = 12.78 and confidence intervals are 5.60 and 19.95. Methodologists are now increasingly recommending that researchers should report the precise P value. For example, P = 0.023 rather than P < 0.05 [10]. Further, to use P = 0.05 "is an anachronism. It was settled on when P values were hard to compute and so some specific values needed to be provided in tables. Now calculating exact P values is easy (i.e., the computer does it) and so the investigator can report (P = 0.04) and leave it to the reader to (determine its significance)"[11].

Hypothesis Tests

A statistical test provides a mechanism for making quantitative decisions about a process or processes. The purpose is to make inferences about population parameter by analyzing differences between observed sample statistic and the results one expects to obtain if some underlying assumption is true. This comparison may be a single observed value versus some hypothesized quantity or it may be between two or more related or unrelated groups. The choice of statistical test depends on the nature of the data and the study design.

Neyman and Pearson proposed this process to circumvent Fisher's subjective practice of assessing strength of evidence against the null effect. In its usual form, two hypotheses are put forward: a null hypothesis (usually a statement of null effect) and an alternative hypothesis (usually the opposite of null hypothesis). Based on the outcome of the hypothesis test one hypothesis is rejected and accept the other based on a previously predetermined arbitrary benchmark. This bench mark is designated the P value. However, one runs into making an error: one may reject one hypothesis when in fact it should be accepted and vise versa. There is type I error or *á* error (i.e., there was no difference but really there was) and type II error or *â* error (i.e., when there was difference when actually there was none). In its simple format, testing hypothesis involves the following steps:

1. Identify null and alternative hypotheses.
2. Determine the appropriate test statistic and its distribution under the assumption that the null hypothesis is true.
3. Specify the significance level and determine the corresponding critical value of the test statistic under the assumption that null hypothesis is true.
4. Calculate the test statistic from the data.

Having discussed P value and hypothesis testing, fallacies of hypothesis testing and P value are now looked into.

Fallacies of Hypothesis Testing

In a paper I submitted for publication in one of the widely read medical journals in Nigeria, one of the reviewers commented on the age-sex distribution of the participants, "Is there any difference in sex distribution, subject to chi square statistics"? Statistically, this question does not convey any query and this is one of many instances among medical researchers (postgraduate supervisors alike) in which test of hypothesis is quickly and spontaneously resorted to without due consideration to its appropriate application. The aim of my research was to determine the prevalence of diabetes mellitus in a rural community; it was not part of my objectives to determine any association between sex and prevalence

of diabetes mellitus. To the inexperienced, this comment will definitely prompt conducting test of hypothesis simply to satisfy the editor and reviewer such that the article will sail through. However, the results of such statistical tests becomes difficult to understand and interprete in the light of the data. (The result of study turned out that all those with elevated fasting blood glucose are females). There are several fallacies associated with hypothesis testing. Below is a small list that will help avoid these fallacies.

1. Failure to reject null hypothesis leads to its acceptance. (**No.** When you fail to reject null hypothesis it means there is insufficient evidence to reject)
2. The use of *á* = 0.05 is a standard with an objective basis (**No.** *á* = 0.05 is merely a convention that evolved from the practice of R.A. Fisher. There is no sharp distinction between "significant" and "not significant" results, only increasing strong evidence against null hypothesis as P becomes smaller. (P=0.02 is stronger than P=0.04)
3. Small P value indicates large effects (**No.** P value does not tell anything about size of an effect)
4. Statistical significance implies clinical importance. (**No.** Statistical significance says very little about the clinical importance of relation. There is a big gulf of difference between statistical significance and clinical significance. By statistical definition at á = 0.05, it means that 1 in 20 comparisons in which null hypothesis is true will result in P < 0.05!. Finally, with these and many fallacies of hypothesis testing, it is rather sad to read in journals how significance testing has become an insignificance testing.

Fallacies of P Value

Just as test of hypothesis is associated with some fallacies so also is P value with common root causes, " It comes to be seen as natural that any finding worth its salt should have a P value less than 0.05 flashing like a divinely appointed stamp of approval"[12]. The inherent subjectivity of Fisher's P value approach and the subsequent poor understanding of this approach by the medical community could be the reason why P value is associated with myriad of fallacies. Thirdly, P value produced by researchers as mere "passports to publication" aggravated the situation [13]. We were earlier on awakened to the inadequacy of the P value in clinical trials by Feinstein [14],

"The method of making statistical decisions about 'significance' creates one of the most devastating ironies in modern biologic science. To avoid usual categorical data, a critical investigator will usually go to enormous efforts in mensuration. He will get special machines and elaborate technologic devices to supplement his old categorical statement with new measurements of 'continuous' dimensional data. After all this work in getting 'continuous' data, however, and after calculating all the statistical tests of the data, the investigator then makes the final decision about his results on the basis of a completely arbitrary pair of dichotomous categories. These categories, which are called 'significant' and 'nonsignificant', are usually demarcated by a P value of either 0.05 or 0.01, chosen according to the capricious dictates of the statistician, the editor, the reviewer or the granting agency. If the level demanded for 'significant' is 0.05 or lower and the P value that emerge is 0.06, the investigator may be ready to discard a well-designed, excellently conducted, thoughtfully analyzed, and scientifically important experiment because it failed to cross the Procrustean boundary demanded for statistical approbation.

We should try to understand that Fisher wanted to have an index of measurement that will help him to decide the strength of evidence against null effect. But as it has been said earlier his idea was poorly understood and criticized and led to Neyman and Pearson to develop hypothesis testing in order to go round the problem. But, this is the result of their attempt: "accept" or "reject" null hypothesis or alternatively "significant" or "non significant". The inadequacy of P value in decision making pervades all epidemiological study design. This head-or-tail approach to test of hypothesis has pushed the stakeholders in the field (statistician, editor, reviewer or granting agency) into an ever increasing confusion and difficulty. It is an accepted fact among statisticians of the inadequacy of P value as a sole standard judgment in the analysis of clinical trials [15]. Just as hypothesis testing is not devoid of caveats so also P values. Some of these are exposed below.

1. The threshold value, P < 0.05 is arbitrary. As has been said earlier, it was the practice of Fisher to assign P the value of 0.05 as a measure of evidence against null effect. One can make the "significant test" more stringent by moving to 0.01 (1%) or less stringent moving the borderline to 0.10 (10%). Dichotomizing P values into "significant" and "non significant" one loses information the same way as demarcating laboratory finding into normal" and "abnormal", one may ask what is the difference between a fasting blood glucose of 25mmol/L and 15mmol/L?

2. Statistically significant (P < 0.05) findings are assumed to result from real treatment effects ignoring the fact that 1 in 20 comparisons of effects in which null hypothesis is true will result in significant finding (P < 0.05). This problem is more serious when several tests of hypothesis involving several variables were carried without using the appropriate statistical test, e.g., ANOVA instead of repeated t-test.

3. Statistical significance result does not translate into clinical importance. A large study can detect a small, clinically unimportant finding.

4. Chance is rarely the most important issue. Remember that when conducting a research a questionnaire is usually administered to participants. This questionnaire in most instances collect large amount of information from several variables included in the questionnaire. The manner in which the questions

where asked and manner they were answered are important sources of errors (systematic error) which are difficult to measure.

What Influences P Value?

Generally, these factors influence P value.

1. *Effect size.* It is a usual research objective to detect a difference between two drugs, procedures or programmes. Several statistics are employed to measure the magnitude of effect produced by these interventions. They range: r^2, $ç^2$, $ù^2$, R^2, Q^2, Cohen's d, and Hedge's g. Two problems are encountered: the use of appropriate index for measuring the effect and secondly size of the effect. A 7kg or 10 mmHg difference will have a lower P value (and more likely to be significant) than a 2-kg or 4 mmHg difference.

2. *Size of sample.* The larger the sample the more likely a difference to be detected. Further, a 7 kg difference in a study with 500 participants will give a lower P value than 7 kg difference observed in a study involving 250 participants in each group.

3. *Spread of the data.* The spread of observations in a data set is measured commonly with standard deviation. The bigger the standard deviation, the more the spread of observations and the lower the P value.

P Value and Statistical Significance: An Uncommon Ground

Both the Fisherian and Neyman-Pearson (N-P) schools did not uphold the practice of stating, "P values of less than 0.05 were regarded as statistically significant" or "P-value was 0.02 and therefore there was statistically significant difference." These statements and many similar statements have criss-crossed medical journals and standard textbooks of statistics and provided an uncommon ground for marrying the two schools. This marriage of inconvenience further deepened the confusion and misunderstanding of the Fisherian and Neyman-Pearson schools. The combination of Fisherian and N-P thoughts (as exemplified in the above statements) did not shed light on correct interpretation of statistical test of hypothesis and p-value. The hybrid of the two schools as often read in medical journals and textbooks of statistics makes it as if the two schools were and are compatible as a single coherent method of statistical inference [4, 23, 24]. This confusion, perpetuated by medical journals, textbooks of statistics, reviewers and editors, have almost made it impossible for research report to be published without statements or notations such as, "statistically significant" or "statistically insignificant" or "P<0.05" or "P>0.05".Sterne, then asked "can we get rid of P-values? His answer was "practical experience says no-why? [21]"

However, the next section, "P-value and confidence interval: a common ground" provides one of the possible ways out of the seemingly insoluble problem. Goodman commented on P–value and confidence interval approach in statistical inference and its ability to solve the problem. "The few efforts to eliminate P values from journals in favor of confidence intervals have not generally been successful, indicating that the researchers' need for a measure of evidence remains strong and that they often feel lost without one"[6].

P Value and Confidence Interval: A Common Ground

Thus, so far this paper has examined the historical evolution of 'significance' testing as was initially proposed by R.A. Fisher. Neyman and Pearson were not accustomed to his subjective approach and therefore proposed 'hypothesis testing' involving binary outcomes: "accept" or "reject" null hypothesis. This, as we saw did not "solve" the problem completely. Thus, a common ground was needed and the combination of P value and confidence intervals provided the much needed common ground.

Before proceeding, we should briefly understand what confidence intervals (CIs) means having gone through what p-values and hypothesis testing mean. Suppose that we have two diets A and B given to two groups of malnourished children. An 8-kg increase in body weight was observed among children on diet A while a 3-kg increase in body weights was observed on diet B. The effect in weight increase is therefore 5kg on average. But it is obvious that the increase might be less than 3kg and also more than 8kg, thus a range can be represented and the chance associated with this range under the confidence intervals. Thus, for 95% confidence interval in this example will mean that if the study is repeated 100 times, 95 out of 100 the times, the CI contain the true increase in weight. Formally, 95% CI: "the interval computed from the sample data which when the study is repeated multiple times would contain the true effect 95% of the time."

In the 1980s, a number of British statisticians tried to promote the use of this common ground approach in presenting statistical analysis[16, 17, 18]. They encouraged the combine presentation of P value and confidence intervals. The use of confidence intervals in addressing hypothesis testing is one of the four popular methods journal editors and eminent statisticians have issued statements supporting its use [19]. In line with this, the American Psychological Association's Board of Scientific Affairs commissioned a white paper, "Task Force on Statistical Inference". The Task Force suggested,

"When reporting inferential statistics (e.g. t - tests, F - tests, and chi-square) include information about the obtained ….. value of the test statistic, the degree of freedom, the probability of obtaining a value as extreme as or more extreme than the one obtained [i.e., the P value]…. Be sure to include sufficient descriptive statistics [e.g. per-cell sample size, means, correlations, standard deviations]…. The reporting of confidence intervals [for estimates of parameters, for functions of parameter such as differences in means, and for effect sizes] can be an extremely effective way

of reporting results… because confidence intervals combine information on location and precision and can often be directly used to infer significance levels"[20].

Jonathan Sterne and Davey Smith came up with their suggested guidelines for reporting statistical analysis as shown in the box[21]:

Box 1: Suggested guidance's for the reporting of results of statistical analyses in medical journals.

1. The description of differences as statistically significant is not acceptable.

2. Confidence intervals for the main results should always be included, but 90% rather than 95% levels should be used. Confidence intervals should not be used as a surrogate means of examining significance at the conventional 5% level. Interpretation of confidence intervals should focus on the implication (clinical importance) of the range of values in the interval.

3. When there is a meaningful null hypothesis, the strength of evidence against it should be indexed by the P value. The smaller the P value, the stronger is the evidence.

4. While it is impossible to reduce substantially the amount of data dredging that is carried out, authors should take a very skeptical view of subgroup analyses in clinical trials and observational studies. The strength of the evidence for interaction-that effects really differ between subgroups – should always be presented. Claims made on the basis of subgroup findings should be even more tempered than claims made about main effects.

5. In observational studies it should be remembered that considerations of confounding and bias are at least as important as the issues discussed in this paper.

Since the 1980s when British statisticians championed the use of confidence intervals, journal after journal are issuing statements regarding its use. In an editorial in Clinical Chemistry, it read as follows,

"There is no question that a confidence interval for the difference between two true (i.e., population) means or proportions, based on the observed difference between sample estimate, provides more useful information than a P value, no matter how exact, for the probability that the true difference is zero. The confidence interval reflects the precision of the sample values in terms of their standard deviation and the sample size ….."[22]

On the final note, it is important to know why it is statistically superior to use P value and confidence intervals rather than P value and hypothesis testing:

1. Confidence intervals emphasize the importance of estimation over hypothesis testing. It is more informative to quote the magnitude of the size of effect rather than adopting the significant-nonsignificant hypothesis testing.

2. The width of the CIs provides a measure of the reliability or precision of the estimate.

3. Confidence intervals makes it far easier to determine whether a finding has any substantive (e.g. clinical) importance, as opposed to statistical significance.

4. While statistical significant tests are vulnerable to type I error, CIs are not.

5. Confidence intervals can be used as a significance test. The simple rule is that if 95% CIs does not include the null value (usually zero for difference in means and proportions; one for relative risk and odds ratio) null hypothesis is rejected at 0.05 levels.

6. Finally, the use of CIs promotes cumulative knowledge development by obligating researchers to think meta-analytically about estimation, replication and comparing intervals across studies[25]. For example, in a meta-analysis of trials dealing with intravenous nitrates in acute myocardial infraction found reduction in mortality of somewhere between one quarter and two-thirds. Meanwhile previous six trials [26] showed conflicting results: some trials revealed that it was dangerous to give intravenous nitrates while others revealed that it actually reduced mortality. For the six trials, the odds ratio, 95% CIs and P-values are: OR = 0.33 (CI = 0.09, 1.13, P = 0.08); OR = 0.24 (CI = 0.08, 0.74, P = 0.01); OR = 0.83(CI = 0.33, 2.12, P = 0.07); OR = 2.04 (CI = 0.39, 10.71, P = 0.04); OR = 0.58 (CI = 0.19. 1.65; P = 0.29) and OR = 0.48 (CI = 0.28, 0.82; P = 0.007). The first, third, fourth and fifth studies appear harmful; while the second and the sixth appear useful (in reducing mortality).

What is to be done?

While it is possible to make a change and improve on the practice, however, as Cohen warns, "Don't look for a magic alternative … It does not exist" [27].

1. The foundation for change in this practice should be laid in the foundation of teaching statistics: classroom. The curriculum and class room teaching should clearly differentiate between the two schools. Historical evolution should be clearly explained so also meaning of "statistical significance". The classroom teaching of the correct concepts should begin at undergraduate and move up to graduate classroom instruction, even if it means this teaching would be at introductory level.

2. We should promote and encourage the use of confidence intervals around sample statistics and effect sizes. This duty lies in the hands of statistics

teachers, medical journal editors, reviewers and any granting agency.

3. Generally, researchers, preparing on a study are encouraged to consult a statistician at the initial stage of their study to avoid misinterpreting the P value especially if they are using statistical software for their data analysis.

REFERENCES

1. **Goodman SN**. P value hypothesis and likelihood: implications for epidemiology of a neglected historical debate. Amer Journ Epidemiol 1993; 137: 485-96.
2. **Lehmann El.** The Fisher, Neyman-Pearson theories of testing hypothesis: one theory or two? Journ Amer Stat Assoc 1993; 88:1242 – 9
3. **Goodman SN.** Toward evidence-based medical statistics: the P-value fallacy. Ann intern Med 1999; 130:995-1004.
4. **Berking JA,** Begg CB, Louis JA. An assessment of publication bias using a sample of published clinical trials. Journ Amer Stat Assoc. 1989; 84:38-92.
5. **Easterbrook PJ,** Berking JA, Gopalan R and Mathews DR. Publication bias in clinical research. Lancet 1991. 337:887 – 892
6. **Dickerson K,** Min YI and Meinert CL. Factors influencing publication of research results: follow-up of applications submitted to two institutional review boards. Journ Amer Med Assoc 1992; 263:374 -378
7. **Stern JAC,** Smith GD. Sifting the evidence-what is wrong with significance tests? Br Med Journ 2001; 322:226-231.
8. **Fisher RA.** Statistical methods for research workers. London Oliver and Boyd, 1950:80.
9. **Bakan D.** The test of significance in psychological research. Psychology Bulletin 1960, 66: 423-437
10. **Greenwald AG,** Richard G, Richard 3H and Donal G. Effect sizes and P value: what should be reported and what should be replicated? Psychophysiology 1996:33:175 – 183
11. **Wainer H,** Robinson DH. Shaping of the practice of null hypothesis significance testing. Educational Researcher 2002:32:22 -30.
12. **Jekel JF.** Should we stop using the P value in descriptive studies? Paediatrics 1977; 60:124 – 25.
13. **Mainland D.** Statistical ritual in clinical journals: is there a cure?-I. Br Med Journ 1984; 288:841-43
14. **Feinstein AR.** Clinical biostatistics. St Louis: CV Mosby; 1977: 66-70
15. **Diamond GA,** Forrester JS. Clinical trials and statistical verdicts: probable ground for appeal. Ann Intern Med 1983; 98:385-394
16. **Altman DG,** Gore SM, Gardner MJ and Pocock SJ. Statistical guidelines for contributors to medical journals. Br Med Journ 1983; 286: 1989 – 93.
17. **Gardner MJ,** Altman DG. Confidence intervals rather than P-value: estimation rather than hypothesis testing. Br Med Journ1986; 292:746-750.
18. **Gardner MJ,** Altman DG. Statistics with confidence – confidence intervals and statistical guidance. BMJ Books, London 1989.
19. Iacobucci D. On P value. Editorial. Journal Cons Res. 2005
20. American Psychological Association. Publication Manual, 5th Ed, Washington, DC: American Psychological Association.
21. **Sterne JAC,** Davey Smith. Suggested guidelines for the reporting of results of statistical analysis in medical journals. Br Med Journ 2001; 322:226-231.
22. **Eugene KH.** On P value and confidence intervals (Why can't we P with more confidence). Editorial. Chin Chem 1993; 39(6):927 – 928.
23. **Royal R.** Statistical Evidence: a likelihood primer. Monographs on statistics and applied probability #71. London: Chapman and Hall; 1997.
24. **Gigerenzer G,** Swijtink Z, Porter T, Daston L, Beatty J and Kruger L. The Empire of Chance. Cambridge, UK: Cambridge Univ Pr; 1989.
25. **Thomas B.** What future quantitative social science research could look like: confidence interval for effect sizes. Educ Research 2002; 31:25-32.
26. **Yusuf S,** Collins R, Mac Mahon, Peto R. Effect of intravenous nitrates on mortality in acute myocardial infarction: an overview of randomized trials. Lancet 1988; 1:1088-92
27. **Cohen J.** Things I Have Learned (So Far). Amer Psychologist 1990; 45:997-1003.