Scand J Work Environ Health 1997;23(2):152-114

<https://doi.org/10.5271/sjweh.193>

Issue date: Apr 1997

"Significant-itis" -- an obsession with the P-value

by Chia K-S

The following article refers to this text: [1989;15\(6\):440-442](#)

Key terms: [confidence intervals](#); [hypothesis testing](#); [statistical significance](#)

This article in PubMed: www.ncbi.nlm.nih.gov/pubmed/9167239

This work is licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/).

“Significant-itis” — an obsession with the P-value

by Kee-Seng Chia, MD¹

Chia K-S. “Significant-itis” — an obsession with the P-value. *Scand J Work Environ Health* 1997;23(2):152—4.

Key terms confidence intervals, hypothesis testing, statistical significance.

Figure 1. “Significant-itis” — a maniac-depressive disorder characterized by an obsession with the P-value.

Many researchers perform statistical analyses merely with the aim of obtaining statistically significant results. This obsession with the P-value is characterized by compulsive hypothesis testing of all results first without assessing the role of systematic errors and the necessity for evaluating the role of sampling error. This practice has led to many erroneous conclusions. Young researchers should be taught to be convinced first of the magnitude of the relationship, followed by the identification of systematic errors before assessing sampling error.

Most original research papers are filled with the ubiquitous P-value. In fact it is widely believed that editors are not keen to publish papers if their results are not statistically significant. This belief has driven many researchers to view statistical analysis merely as a means of obtaining a statistically significant result; in the long run, these researchers develop a compulsive obsession

which can be termed “significant-itis” (figure 1). The aim of this commentary is to highlight the limitations of the P-value in order to control this epidemic.

Statistical significance is not as informative as confidence intervals

Tests of significance force the researcher to decide if a result is statistically significant on the basis of the P-value. The 5% level used is purely arbitrary. Rumor has it that the statistician Sir Ronald Fisher was lying in his bathtub contemplating this very perplexing issue. While deep in thought, he pulled the plug accidentally and the water started draining from the tub. As the water level descended, he saw his five toes sticking out. That observation settled the issue; statistical significance is to be set at the 5% level! One might argue that if Fisher was more observant, he would have set it at 10%, unless he was one legged or had a fixed deformity of the other foot!

It is therefore absurd to label a P-value of 0.051 insignificant while that of 0.049 is considered significant. Quoting the exact P-value may be a little more informative, but most tests of significance derive the P-value based on estimation procedures. Hence the exact P-values may be of little meaning. Furthermore, the meaning of the exact P-value is not easily obvious.

Confidence intervals are more informative than the P-value and are easily interpretable. P-values fail to convey the magnitude of the differences between study groups; confidence intervals, on the other hand, provide the range within which the differences are likely to be found. Wherever possible, confidence intervals should be reported rather than P-values (1—3). Methods for calculating confidence intervals are available in most statistical textbooks but are neatly compiled in the *British Medical Journal*'s publication, *Statistics with Confi-*

¹ Centre for Environmental and Occupational Health, Department of Community, Occupational & Family Medicine, National University of Singapore.

Reprint requests to: Dr K-S Chia, Centre for Environmental and Occupational Health, Department of Community, Occupational & Family Medicine, National University of Singapore, Lower Kent Ridge Road, Singapore 119074, Republic of Singapore. [E-mail: cofcks@nus.sg]

dence (3). Most statistical software provide the confidence interval or at least the standard error (SE) of the estimate. Confidence intervals can then be manually computed from the SE. However, one should know the probability distribution used (eg, normal, t- or F-distribution) and the number of degrees of freedom rather than indiscriminately applying the formula: $\pm 1.96 \times \text{SE}$!

A statistically significant result may be clinically irrelevant

With powerful statistical packages on microcomputers, it is very tempting to conduct tests of significance on every single analysis. The researcher can be easily distracted by the P-value rather than by the magnitude of the difference. Consider the example of a study evaluating the presence of a workplace respiratory allergen using the pre- and postshift value for forced expiratory volume in 1 s. A statistically significant difference was found, but the postshift value was only 150 ml/s less than the preshift value. Clearly, the difference could well have been due to measurement error or diurnal variation. However, the statistical significance may suggest to the researcher to conclude that a workplace allergen is present.

As a general guideline, one should not proceed with tests of significance (or even construct confidence intervals) unless the magnitude of the difference is clinically significant.

Statistical significance is dependent on the sample size

Both tests of significance and confidence intervals are derived from the SE. The main difference is that the test of significance uses a summary statistic (eg, t-value, F-value), whereas confidence intervals are computed directly from the SE. The SE is directly related to the variability of the observation and inversely related to the sample size. Hence given similar variability, with a larger sample size, the SE will be smaller. For example, a difference in the mean blood lead levels of 2 groups of workers was originally not statistically significant. If the means and standard deviations are kept constant, a larger sample size will produce statistically significant results.

Some microcomputer statistical packages offer 2 types of unpaired t-tests: a t-test for equal variance and a t-test for unequal variance. To help the user, a test of significance is performed on the variance of the 2 samples. A small difference in the variance may have a statistically significant difference because of a large sample size. Conversely, a large difference in variance may not produce a statistically significant difference if the sample size is small. Hence, one should not rely solely on the test of significance to decide if the variances are equal.

As a further illustration, many researchers identify potential confounders between the study and comparison

groups by performing tests of significance. In a small study, a large difference in the distribution of potential confounders between the study and reference group will not be statistically significant. As a consequence, they will not be controlled for in the subsequent analysis. Conversely, small differences in the distribution of the potential confounders are vigorously adjusted in large studies. This erroneous practice has been previously highlighted and well illustrated by Hernberg (4).

The preceding illustrations further support the point that, since statistical significance is dependent on sample size, the magnitude of the difference should be the main consideration rather than whether the difference is statistically significant, at least in large studies.

Whatever the result of the test of significance, there is always a risk of type I and type II error

Type I error (α error) is defined as rejecting the null hypothesis when one should have not have rejected it. Type II error (β error) is not rejecting the null hypothesis when one should have rejected it. The difference is elegantly described in the following story: "Once upon a time, there was a King who was very jealous of his Queen. He had two knights: handsome Alpha and ugly Beta. The Queen was in love with Beta. The King beheaded Alpha because he suspected him to be having an affair with the Queen. The Queen fled with Beta & lived happily ever after. The King made both types of errors: He suspected a relationship with Alpha when there was none (Type I error). He failed to detect a relationship with Beta when there was one (Type II error)" (5).

When a test of significance shows $P < 0.05$, there is still a less than 5% chance that the "truth" is that there is no difference. Similarly, when $P > 0.05$, there is a possibility that there is a difference, but, because of the small sample size, it was not detected. Hence, whatever the result of the test of significance, there is always a risk of sampling error.

Even if the result is statistically significant, there are other sources of errors

"Before you analyze the data you have, you must remember where you got it from". This advice by Sir Austin Bradford Hill underscores the limitations of statistical analysis. Tests of significance and confidence intervals merely assess the magnitude of sampling error. There is at least another major source of error, that of systematic error. It is unfortunate that most statistical courses focus on sampling errors, whereas courses on epidemiologic methods cover systematic errors.

Systematic errors are not easily controlled. These errors include biases and confounding. Biases are related to the quality of the data. Biases have to be meticulously avoided at the design and data collection stage. Con-

founding, on the other hand, can be managed at both the design and data analysis stage. There are excellent accounts on systematic errors in major epidemiologic textbooks (6, 7).

Concluding remarks

The aim of statistical analysis is to make some sense out of the mass of raw data collected. Unfortunately, most researchers indiscriminately apply tests of significance or even construct confidence intervals without first observing any relationship in the data. A further pitfall is to draw conclusions without any evaluation for systematic errors. Perhaps a more intuitive approach would be to first identify the relationship, followed by evaluating for systematic errors, and finally assessing the random errors.

References

1. Rothman K. A show of confidence. *N Engl J Med* 1978;299:1362—3.
2. Poole C, Lanes S, Rothman K. Analysing data from ordered categories. *N Engl J Med* 1984;311:1382.
3. Gardner MJ, Altman DG, editors. *Statistics with confidence*. London: British Medical Journal, 1989.
4. Hernberg S. Significance testing of potential confounders and other properties of study groups — misuse of statistics. *Scand J Work Environ Health* 1996;22:315—6.
5. Wasseldhiel-Smoller S. The scientific method. In: Wasseldhiel-Smoller S. *Biostatistics and epidemiology: a primer for health professionals*, section 1. Springer-Walleg, 1990:12.
6. Rothman K. Objectives of epidemiologic study designs. In: Rothman K, editor. *Modern epidemiology*. Boston (MA): Little, Brown and Company, 1986:89—94.
7. Hernberg S. Internal validity, precision and generalization. In: Hernberg S. *Introduction to occupational epidemiology*. Chelsea (MI): Lewis Publishers, 1992:115—28.