Use and Misuse of p-Values in Designed and Observational Studies: Guide for Researchers and Reviewers

David A. Ludwig

Ludwig DA. ***Use and misuse of p-values in designed and observational studies: guide for researchers and reviewers.*** **Aviat Space Environ Med 2005; 76:675–80.**

Analysis of scientific data involves many components, one of which is often statistical testing with the calculation of p-values. However, researchers too often pepper their papers with p-values in the absence of critical thinking about their results. In fact, statistical tests in their various forms address just one question: does an observed difference exceed that which might reasonably be expected solely as a result of sampling error and/or random allocation of experimental material? Such tests are best applied to the results of designed studies with reasonable control of experimental error and sampling error, as well as acquisition of a sufficient sample size. Nevertheless, attributing an observed difference to a specific treatment effect requires critical thinking on the part of the scientist. Observational studies involve data sets whose size is usually a matter of convenience with results that reflect a number of potentially confounding factors. In this situation, statistical testing is not appropriate and p-values may be misleading; other more modern statistical tools should be used instead, including graphic analysis, computer-intensive methods, regression trees, and other procedures broadly classified as bioinformatics, data mining, and exploratory data analysis. In this review, the utility of p-values calculated from designed experiments and observational studies are discussed, leading to the formation of a decision tree to aid researchers and reviewers in understanding both the benefits and limitations of statistical testing.

Keywords: statistics, inference, p-values, observational studies, experimental design.

DESPITE A MULTITUDE of warnings and explanations from all disciplines of science, researchers often misuse and misinterpret the results of statistical tests, nor can they correctly explain the meaning of a p-value (3,6). Research articles are often littered with p-values in the absence of appropriate critical analysis. This problem is all too familiar to the author, who is the consulting statistician for *Aviation, Space, and Environmental Medicine*. Throughout academia, industry, and government research laboratories, misconceptions regarding statistical testing lead authors to ask "Is it significant?" or say, "I want to know if there is a real difference." In this review, I shall provide for the journal's contributors and readers a clear explanation of the true nature of statistical testing and provide a flowchart to aid in evaluating the results of statistical analyses.

Examples of Misunderstanding

In a recent paper, I stated that the observed mean for the treatment group was 10 units higher than that for the control group. A reviewer complained (in bold type, no less) that I could not make such a statement unless the difference was statistically significant. Alas, many scientists seem to believe that the observed results of an experiment or observational study are not factual and, therefore, cannot be discussed unless some type of statistical sanctification is invoked.

While the observed means or the mean difference in my study may not be factual as a result of sampling variability or measurement error, it is certain ($p = 1.0$!) that there was a 10-unit difference between groups. No author should be faulted (and in most cases should be praised) for discussing the actual outcome of the experiment; doing so does not require a statistical test nor is the result probabilistic. For a specific experiment, the results occur with absolute certainty.

A second example of this kind of misconception concerns an article that I reviewed that reported commercial airline accident rates over a span of 10 yr. Such accidents are subject to exact enumeration and thorough investigation, so we can assume that the database included every accident during that period. The author presented the data with 95% confidence limits and an index of sampling error, despite the fact that this was the entire population and not a sample! When asked to remove the confidence limits, the author suggested that other kinds of errors (i.e., incorrect reporting or computer entry mistakes) could have influenced the rates, and I had to point out that such non-sampling errors are not subject to statistical analysis unless repeated mea-

From the Dept. of Pediatrics, Medical College of Georgia, Georgia Prevention Institute, Augusta, GA.

This manuscript was received for review in February 2005. It was accepted for publication in April 2005.

Address reprint requests to: Professor David A. Ludwig, Ph.D., Dept. of Pediatrics, Medical College of Georgia, Georgia Prevention Institute, MCG Annex, H.S. 1640, Augusta, GA 30912-3710; dludwig@mcg.edu.

Dr. Ludwig is the current statistical consultant to ASEM. He has more than 25 years of experience teaching graduate level statistics and is a long-time statistical consultant to NASA, the U.S. Air Force, and the U.S. Army. A former Fulbright Scholar in biostatistics, he is currently a Professor of Pediatrics and Biostatistics at the Medical College of Georgia and the Georgia Prevention Institute.

sures (i.e., reliability checks) are performed, which was not the case here.

As a statistician with long experience in human experimentation and observational studies, I believe that these examples reflect a lamentably common state of misinformation among medical scientists. The remainder of this paper will be devoted to helping readers understand the correct use and limitations of statistical inference.

Statistical Tests

Ask research scientists why they wish to perform statistical tests and the answer is likely to be either, "I want to know whether there is a difference" or "I want to see if the difference is significant." Unfortunately, statistical tests cannot answer either question. Establishing a difference is a matter of looking at the numbers: does the mean of the treatment group equal the mean of the control group or does it differ? Evaluating significance is more complex: the scientist likely wants to know whether the results are important, notable, consequential, vital, crucial, serious, critical, momentous, and/or weighty, but statistical tests do not address those issues. Even the more limited question of "statistical significance" is vague because "significance" has many meanings and interpretations. In fact, statistical testing in its various forms addresses just one question: does the observed difference exceed that which might reasonably be expected solely as a result of sampling error and/or random allocation of experimental material? If the answer is "no," the difference could be a chance occurrence, simply a function of the randomization or sampling process.

In a designed experiment, material is assigned to experimental conditions, the treatments are applied, and results are then observed. Even if the treatment has no effect whatever, the observed means for different groups will likely differ due to the effects of experimental error and sampling error. Experimental error should be a minor concern if the experiment is well designed with tight controls and good measurement techniques. Sampling error reflects the fact that different initial random allocations will produce slightly different results even when the treatment has no effect. Statistical tests compare the specific observation (experimental outcome or difference) to that which might be expected if there were no treatment effect, i.e., sampling variation was the only factor operating in the experiment.

Since there are many possible outcomes to the rerandomization of an experiment, the observed results are compared with a distribution of outcomes derived solely as a result of sampling error. These "reference distributions," which reflect the differences that might be observed in multiple repetitions of a noise-only experiment, can be found in the back of any statistics or biostatistics text (e.g., z, t, F, χ^2). Statistical tests and their resulting p-values are generated when the difference observed in the experiment is compared with the reference distribution of differences in the absence of any treatment effect. If the observed result greatly exceeds that which might be expected solely as a result of random sampling or random allocation, then one may conclude that some other effect influenced the results. That is, something is at work that influenced the means beyond the effect of only sampling error. Note that the researcher cannot say that there was a treatment effect, only that the result exceeded what might be logically expected due to sampling error alone. Attributing the result to the treatment is a completely different problem that is not addressed by statistical testing, a distinction that is often misunderstood by researchers and is a major cause of the misuse of statistical testing.

Test statistics are merely standardizations of the observed result after considering sample size and observed variation. Large test statistics result when effects are large (big observed differences), experimental error is small (low noise), and/or sample sizes are large (little sampling error). The size of the difference between the observed result and what might be expected solely as a result of sampling error is quantified by where the observed test statistic falls on the sampling-error-only reference distribution (t, F, etc.). If the test statistic is near the center of the distribution, we may conclude that the value of the test statistic is not unlike one that might be obtained if this were a noise-only system. It may then be concluded that there is insufficient evidence to discount sampling variation as a possible reason for the observed difference. If the test statistic is large and falls on the tails of the reference distribution, this is evidence against the "noise-only" hypothesis and it may be concluded that other influences are operating within the system. Depending on the situation, these "other influences" could be anything from the treatment effect to the fact that the researcher was unethical and doctored the numbers. Statistical tests never reveal the source of the observed effect, only whether the latter is rare compared with a noise-only system.

The location of the test statistic is quantified by calculating the percentile of the statistic with regard to the reference distribution. This resembles the way a teacher references an individual's test score to a norm distribution: if a student is in the 90th percentile, we know that approximately 10% (one minus the percentile) of scores exceeds the score obtained by this individual, who can, therefore, be viewed as more the exception than the rule. Similarly, the p-value associated with the test statistic indicates the rarity of the observed test statistic when referenced to a distribution of outcomes in which sampling variation is the only effect. The p-value is one minus the percentile, and indicates the proportion of the reference distribution which is equal to or greater than the observed test statistic. Thus, large test statistics will be associated with low p-values; a p-value of 0.43 indicates a rather common occurrence when referenced to sampling variation alone, while a p-value of 0.01 indicates a rather rare occurrence. That is all we know; we need not sensationalize the results with superlatives such as significant, momentous, or spectacular. Nor do we need to garnish the results with stars, crosses, double asterisks, or smiley faces.

Designed (Randomized) Experiments

Suppose a study meets all the requirements of an ideal experiment including: random assignment of sub-

jects or experimental materials to treatment groups; replication of experimental material within each treatment group; close attention to issues such as measurement, masking, and laboratory procedures; and elimination of potential confounding elements by treating all experimental material alike in every respect except for the independent variable being manipulated. The experiment then yields a 15-unit difference between Treatment Group A and Treatment Group B. The data all seem reasonable and there is no reason to believe that the researcher or laboratory technicians were dishonest in any way. Under these conditions, there are only two possible explanations for the observed difference: one is that a treatment effect was involved and the other is that the treatment had no effect and the result arose purely as a consequence of sampling error (4,7).

Now suppose we construct a test statistic (*t*-test) to determine whether 15 is greater than zero in the context of sampling error. It is a fact that 15 is not equal to zero, but we want to know whether 15 is markedly different from what might be expected if the treatment had no effect, i.e., the only thing operating in this experiment is sampling variation. The test statistic is calculated (correctly) and the associated p-value is 0.03. Can we conclude that the treatment was effective? No! Such a conclusion requires a clinical judgment of efficacy. Can we conclude that the result was significant? No! Significance is a relative term meaning different things to different people over varied circumstances. Can we conclude that the treatment produced a 15-unit change beyond sampling variation? No! Since the effect of sampling variation and the effect of the treatment are mixed together, the 15-unit change cannot all be attributed to the treatment. The only conclusion we can draw is that the observed difference is somewhat unlikely in the absence of a treatment effect, i.e., there was "some" treatment effect. That is it! The actual effect of the treatment may be 3 units or 5 units or 0.006 units!

So, it cannot be concluded that the treatment effect was 15 units, but it might be concluded that the treatment effect was not zero. Notice that it cannot even be said that it is non-zero with certainty, since the end result of the statistical test is a probability. The researcher may feel that 0.03 is sufficiently robust to support a statement that there was some effect of the treatment despite the risk that this is in fact one of those rare situations where the observed 15-unit difference resulted purely from sampling variation. Since this is a properly designed experiment, the researcher might logically attribute this effect to the treatment.

If the p-value = 0.34, it indicates that the experimental result was not markedly different than what might be expected from sampling variability alone. Does that mean there was no treatment effect? No! In the absence of an unusually large test statistic (low p-value), the researcher cannot discount the "noise-only" hypothesis. Does this mean there is no treatment effect? Of course not! There are still two hypotheses left with equally viable explanations for the observed difference: treatment effect and sampling error. The inability to discount the noise-only hypothesis does not mean the treatment had no effect. A test statistic will always yield a large p-value if the sample size is inadequate with respect to the individual variability of the experimental material. If we find a dead animal and know that the cause of death was either poison or old age, but we cannot determine the animal's age, it does not mean the animal was poisoned!

Large p-values will also occur in poorly run experiments using unreliable measurements. Since experimental error and measurement error increases the observed variability between individual pieces of experimental material, sampling error appears larger than it actually is. Treatment effects in excess of sampling variability are more difficult to detect when experimental and/or measurement error are high. In that situation, the experiment was not adequately designed to detect treatment effects in the presence of the great sampling variability. That is why large p-values cannot be used to justify "no effect." The effect may very well be present, but the design was simply inadequate to detect it.

Because a sufficiently large sample allows detection of even the smallest effect no matter how large the experimental error, some investigators would like to scrap the whole process of statistical hypothesis testing (2). However, the ability to detect small differences using a large sample is a virtue of statistical inference. The theory that supports statistical inference is exquisite, but its practice is another matter. Effects come in all sizes. Small effects are just as "real" as large ones. Although separating small effects from sampling variability is difficult, the tools of statistical inference would be of little use if they could detect only large effects. The problem lies in the interpretation of the statistical results. Low p-values say nothing about the clinical relevance of the result or the size of the effect. Small p-values can always be obtained with large samples no matter how much random experimental error is present. In this situation the more important question is one of informed judgment of clinical or practical significance.

Observational Studies

Observational studies usually involve data that is acquired rather than sampled. Studies of this type have a number of names including analytical surveys, correlational studies, and pseudo-experiments. In general, there are two key differences between observational studies and designed experiments. First, there is no form of sampling or random allocation. Observations are often ones of convenience and are not selected or assigned using any type of probabilistic notion. Second, the investigator has minimal control over error variance (experimental error in a true experiment). All types of error, both random and systematic, abound in such studies (5,7,9). Systematic errors create confounds which are typically difficult to adjust for in observational settings.

In a typical observational study, the researcher may believe a certain risk factor is associated with a specific outcome and obtains data that might show such an association. What information might a p-value provide in such a study? Some experts argue adamantly that

statistical testing is inappropriate because such studies lack randomization or random sampling, which are the key ingredients that make inferential statistics work. Without them, the databased estimates that are eventually used to construct statistical tests are too easily biased. Furthermore, the central limit theorem (the basis of all statistical tests and confidence intervals) does not work without randomization or random sampling.

If a statistical test on observational data produces a large p-value (e.g., 0.36), we may be tempted to say there is no evidence that the data is a result of something other than random allocation (sampling error). Although we do not have a random sample or random allocation, the data do not seem to depart much from what we might expect due to sampling variation alone. However, caution is required because an observational study does not follow the rules that allow for this type of conclusion. Since the data are most likely a "grab set," estimates of means and variances that are used in constructing statistical tests may be heavily biased. Statistical tests in designed studies attempt to answer the question, "Given random sampling, what are the chances of this result?" In an observational study we can only ask, "Given the data (acquired without randomization), what are the chances that it is random?" These are two very different chances! In the first case (random sampling), we are aware of the process by which the data was obtained and wish to make statements about the outcome. In the second case (observational data), we have the outcome and are attempting to make statements regarding the process! Traditional statistical tests ask, "Given random sampling or allocation, what are the chances of the data?" This can be written as a statement of conditional probability, p(DATA | RANDOM), (the symbol "|" means, given). This question cannot be answered without random sampling or random allocation. When we perform statistical tests on observational data, the only question we may attempt to answer is p(RANDOM | DATA). Two problems! Traditional statistical tests do not address this question—they only address p(DATA | RANDOM), and without randomization, p(DATA | RANDOM) has no meaning. Most would not confuse p(DEATH | HANGING) and p(HANGING | DEATH)! In the first case, we know the process (HANGING) and wish to make statements regarding the outcome (DEATH). In the second case, we know the outcome (DEATH) and wish to make statements regarding the process (HANGING). We might also ask, "Why are we attempting to answer a question to which we already know the answer?" It is a known fact; data which are "acquired" cannot be random! Researchers must recognize that statistical tests performed on observational data have different meanings from those performed on randomized experiments!

Supposing the p-value for an observational study is small, e.g., $p = 0.03$. Can it be concluded that the independent variable had an effect on the dependent variable beyond that due to sampling variation? No! Since there was no random sampling or allocation, there are many reasons, other than the effect of the independent variable, for data that yield results in excess of sampling variation (4,7). Although a low p-value indicates order that is not random in the data, we cannot attribute this order to any treatment or condition. Perhaps the nonrandom result is purely a function of the nonrandom procedure by which the data were obtained (a statistical Catch-22)! With any grab-set, data may be affected by innumerable confounding variables and biases are in many cases impossible to separate from the effects of the independent variable of interest.

Finally, the most compelling argument against performing statistical tests in an observational study may be that the size of the data set (N) is usually arbitrary, a matter of convenience. Since the p-value is a function of N, this situation in turn generates an arbitrary p-value. If the data set happens to be large, then the p-value will be small and visa versa.

Deduction vs. Induction

In well-designed, randomized experiments the decision process is deductive in nature. Since there are only two possible reasons for the observed outcome, if one can be discounted, then we can deduce the other. Of course, it is not pure deduction since the procedure is probabilistic, but it is close.

Observational studies do not provide such a definitive decision tree. Statistical tests performed on observational data have little meaning and in most cases are detrimental. Acting as though there was an experiment does not create one. Observational studies require hard thought and induction; they are very difficult to do well and cannot be reliably analyzed with mindless computer packages that spew out page after page of p-values. The validity of the results can be determined only through replication or cross-validation. Researchers must also realize that some, perhaps many, observational studies are unanalyzable because there are so many problems and confounds. These confounds arise from the fact that in most observational studies, the data was collected for other purposes than those for which it is now being analyzed. In that case, calculating a p-value will not solve the problem and it may be necessary to scrap the data, painful as that may be.

The Psychology of $p < 0.05$

Researchers and statisticians who perform statistical tests on observational studies are really practicing psychology rather than statistics (13). There appears to be a perception by many scientists that the p-value can right all wrongs. If they obtain $p < 0.05$, they believe that the study is justified and consequently publishable. Psychologically, it makes them feel good!

This is not to condemn observational studies, which can provide important scientific insights and form the majority of articles in this journal. However, the authors should not rely on p-values to make their case. Many other, more appropriate statistical tools are available, including graphic analysis, computer-intensive methods, regression trees, and a variety of other procedures broadly classified as bioinformatics, data mining, and exploratory data analysis (6,12). In the case of observational studies, these modern techniques are much preferred over the experimentally based Neyman-Pearson accept/reject convention of 1928 (11), a convention that is

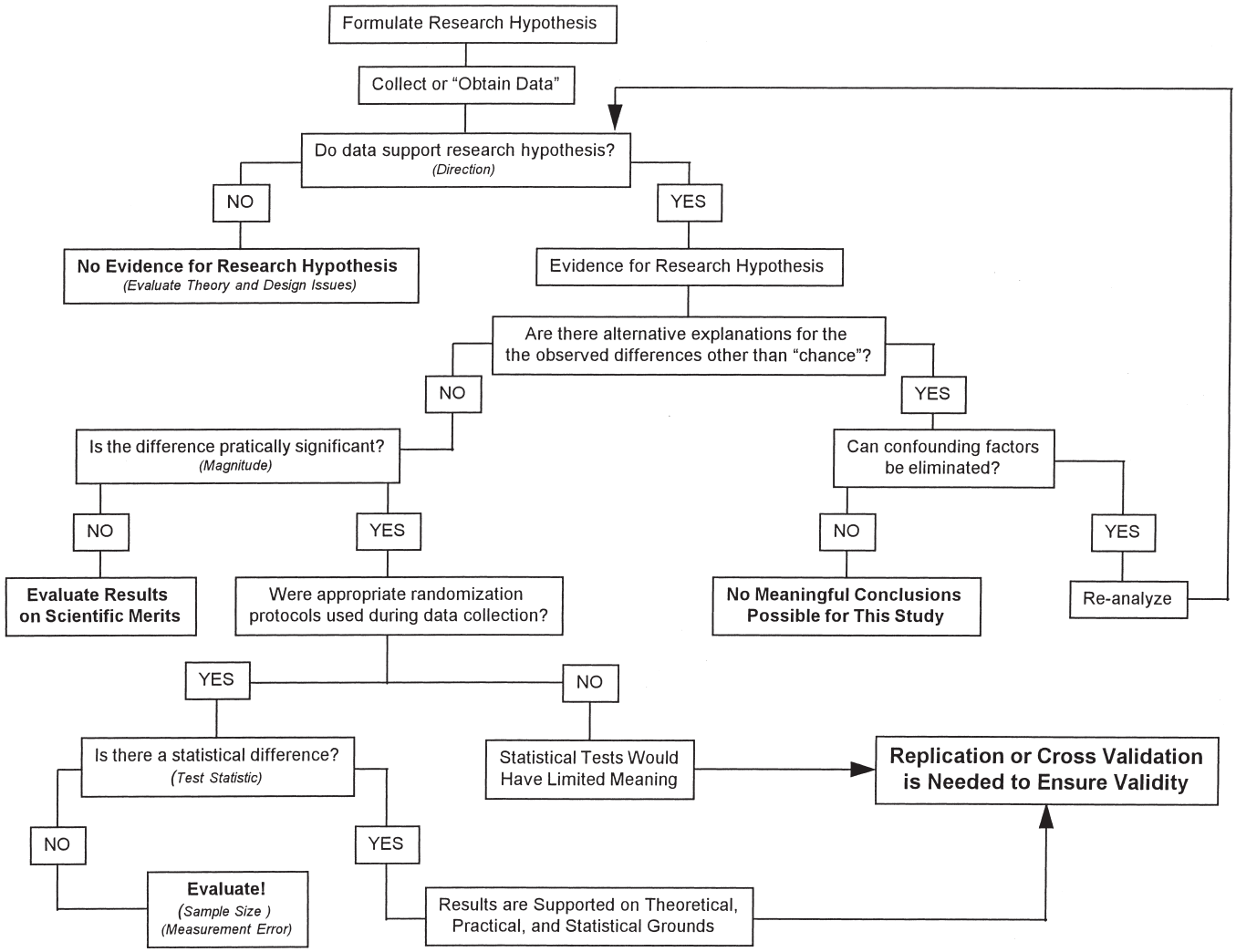

Fig. 1. Decision tree for the analysis and evaluation of designed experiments and observational studies.

completely misunderstood and only applies to the highly specific univariate situations of true experiments (10).

Decision Tree

Fig. 1 presents a flow chart that can be used to evaluate research articles and/or formulate an analysis plan. It shows the major points of this exposition. Note that because statistical tests are meaningless in the face of poor design, evaluation of statistical tests does not appear until late in the decision tree.

Summary and Concluding Statement

To summarize, even in a designed experiment, statistical tests and p-values give very little information because they can answer only the one very specific question. Many other important questions need to be addressed and answers found before valid conclusions can be drawn. In the end, it is design issues that determine the validity of a study (1). In fact, a good practice for authors and reviewers is to read the article or manuscript without looking at or evaluating p-values and then decide if there is sufficient information to support the author's conclusions. If the article or manuscript displays a plethora of p-values, with little or no objective results such as means and effect sizes, it is questionable as to the scientific merits of the study. A good practice is to ask the question "Has the author made a compelling case for the effect, sans p-values?" A good strategy for reviewers would be to initially go thorough the manuscript and black out any p-values or related statements such as "statistically significant." Then read the paper and determine if the results and the write-up provide sufficient evidence to support the research hypothesis. Similarly, authors might first construct their results section without any p-values. Results of statistical tests can then be added to supplement the findings.

Nothing in this paper is new or original. The cited references are only a few from a very long list; many of the references within each article are also very good and highly recommended reading. The misuse of statistical tests abounds despite enlightened scientists' pleas for reform over many decades. Statistical testing makes researchers feel good and are, therefore, hard to give up. Like a drug habit, a kind of dependency on p-values persists and when attempts are made to relegate the p-value to its proper place in science and logical thought, statistical withdrawal occurs! A much clearer

perspective is obtained once researchers and reviewers get themselves out from under the veil of p-values.

An author may say, "I cannot get my paper published unless $p < 0.05$!" Nonsense! The journal *Science* is arguably the premiere scientific journal in the world, yet it publishes few p-values. Watson and Crick discovered the double helix without any p-values (14). Every year pharmaceuticals are recalled despite the fact that the clinical trials that documented their safety and efficacy did so based on a p-value of less than 0.05. An editorial in *Nature Genetics* (a sub-journal of *Nature*) laments that, although hundreds of genetic-association studies have initially found effects in which p was less than 0.05, virtually none of these results have held up under attempted replication (8). This is not to suggest that there is any flaw or incorrect statistical theory associated with statistical testing. The problem is that researchers, reviewers, and editors tend to view p-values as conclusive despite the fact that they address only a specific, limited scientific concept that is meaningful only under very specific conditions.

ACKNOWLEDGMENTS

The author would like to thank Sarah Nunneley for her insightful editing.

REFERENCES

1. Abelson RP. Statistics as principled argument. New Jersey: Lawrence Erlbaum; 1995.
2. Bower B. Null science psychology's statistical status quo draws fire. Science News 1997; 151:356–7.
3. Carver RP. The case against statistical significance testing, revisited. Journal of Experimental Education 1993; 61:287–94.
4. Carver RP. The case against statistical significance testing. Harvard Educational Review 1978; 48:378–99.
5. Cochran WG. Planning and analysis of observational studies. New York: John Wiley; 1983.
6. Cohen J. The earth is round ($p < .05$). American Psychologist 1994; 49:997–1003.
7. Cook TD, Campbell DT. Quasi experimentation: design and analysis issues for field settings. Chicago: Rand McNally; 1979.
8. Editorial. Freely associating. Nature Genetics 1999; 22:1–2.
9. Feinstein AR, Horwitz RI. Double standards, scientific methods, and epidemiologic research. N Engl J Med 1982; 307:1611–7.
10. Goodman SN. P values, hypothesis tests, and likelihood: implications for epidemiology of a neglected historical debate. Am J Epidemiol 1993; 137:485–95.
11. Neyman J, Pearson E. On the use and interpretation of certain test criteria for purposes of statistical inference. Biometrika 1928; 28:175–240.
12. Tukey JW. Data-based graphics: visual display in the decades to come. Statistical Science 1990; 5:327–39.
13. Wang C. Sense and nonsense of statistical inference: controversy, misuse, and subtlety. New York: Marcel Dekker; 1993.
14. Watson JD, Crick FH. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. Nature 1953; 171:737–8.