Special Article

What Your Statistician Never Told You about *P*-Values

Jeffrey Blume, Ph.D., and Jeffrey F. Peipert, M.D., MPH

Abstract   ***(J Am Assoc Gynecol Laparosc 10(4):439–444, 2003)***

We provide a nontechnical overview of what P-values are and what they are not. To determine how P-values ought to be used, reported, and interpreted, we must first clarify the often-overlooked differences between, and proper usages of, significance testing and hypothesis testing. Several clinical examples are given to illustrate these differences, and failure to distinguish between them is seen to be problematic. Common misinterpretations of P-values are explained. Confidence intervals provide essential information where P-values are deficient in doing so and they therefore play an essential role in reporting and interpreting study results.

Has this ever happened to you? After completing an interesting clinical study, you meet with a statistician to review the data analysis. Among other things, he tells you that the *P*-value for your primary comparison is very small, say 0.004. "Great," you say with some confidence, "so the likelihood that my findings are due to chance is 0.004." Not really," says the statistician without even looking up from his printouts. "Well you mean that there is a 99.996% chance that there is a real difference between the study groups or a 0.4% chance that the null hypothesis is true. Right?" "Nope," says the statistician, again without looking up. So you try a different tack: "In the recent Jones et al study they reported a *P*-value of only 0.03 for the same comparison, so at least I have found stronger evidence of the effect. Correct?" "Not exactly," he says, "that study was much larger, so they probably had more evidence of an effect than you do, even though your *P*-value is smaller. I really can't be sure without looking at their results." "But their *P*-value is larger," you exclaim. "Yes, I know," he says after a big sigh, "but *P*-values are weird like that. You can't compare their magnitude unless the sample sizes are equal." You leave the meeting feeling confused, frustrated, and disappointed.

Exchanges like this are unfortunately commonplace. The *P*-value is the most commonly used and, perhaps, most misunderstood statistical concept in clinical research. To the researcher, it is critical; it measures the strength of evidence and provides a means of communicating study results quickly and objectively. To the statistician, it is something of a sore spot; its proper interpretation and use are widely misunderstood, and its adequacy as such a measure is still subject to debate.[1–4] The resulting tension between researchers and statisticians surrounding the proper use and interpretation of *P*-values is understandable, but it is also avoidable.

What is often left unsaid in introductory textbooks and courses on statistics is that the discipline of statistics is itself conflicted about the proper use and interpretation of *P*-values.[4–7] This debate has dragged on since the 1930s without a resolution in sight, despite numerous attempts to resolve the conflict.[3,4,8] As a result, many statisticians simply shun or ignore the issue, sensitive to the fact that the discipline has not yet come to agreement. In fact, several articles advocated replacing *P*-values with effect sizes and confidence intervals (CIs) altogether.[9–13]

We believe a nontechnical roadmap of what *P*-values are and what they are not will help clinicians, researchers, scientists, and statisticians alike. Guidelines will enable authors and readers to interpret, evaluate, and communicate these values in everyday research, while avoiding common pitfalls. To help in the process, we illustrate with two examples from the clinical research setting.

Clinical Illustrations

Our first illustration is a small clinical trial evaluating a new analgesic for chronic pelvic pain. Sixty women are enrolled in the placebo arm and 60 in the experimental therapy arm. Thirty percent of women in the treatment arm experience marked improvement of pain, compared with only 15% of women in the placebo arm. Thus we have an estimated relative risk of 2.0 for the improvement of pain. The investigators report that the *P*-value under the null

hypothesis that the regimens are the same is 0.079 (Fisher's exact test). They interpret this as marginal evidence that the analgesic works better than placebo.

The second example concerns a larger study evaluating a new technique for intrauterine insemination. Two thousand women are recruited and randomized to either the new technique or standard one. Implantation rates are 55% and 50%, respectively. The investigators report that, at the 5% significance level, they rejected the null hypothesis that the two techniques have equal implantation rates. Based on these data alone, they recommend that the new technique be adopted.

Notice the difference in reporting results in the two illustrations. Both are correct, but the goals are different. In the first example the investigators are trying to communicate the strength of evidence in the data by quoting the *P*-value, but in the second example the investigators are reporting how they (and others) should behave. Also, investigators in the second example did not report the strength of evidence in the data (although its magnitude is implied because rejection at the 5% level implies that probability is below 0.05. Indeed, the *P*-value is 0.028). The first is an example of significance testing and the second of hypothesis testing. These two procedures are commonly merged into one grand statistical procedure, causing confusion and misinterpretation of results.

Significance Testing, *P*-Values, and Statistical Evidence

The use of significance tests to interpret the data as statistical evidence was advocated by R. A. Fisher since the early 1920s.[14,15] A significance test is conducted by specifying a null hypothesis, calculating the *P*-value under that hypothesis, and reporting the numerical *P*-value. In clinical research, the null hypothesis would likely state the status quo (there is no difference between treatment groups, the intervention is ineffective etc.). The *P*-value is a tail area probability based on the observed effect; it is calculated as the probability of an effect as large as or larger than the observed effect (more extreme in the tails of the distribution), assuming that the null hypothesis is true.

Fisher claimed that the *P*-value was a measure of the strength of evidence against the null hypothesis and he used it to assess the degree of concordance between data and the null hypothesis.[16] Accordingly, smaller *P*-values indicated stronger evidence against the null hypothesis; the smaller the *P*-value the more inconsistent the data are with the null hypothesis.[17] There was just one caveat: Fisher did not interpret the *P*-value as a probability.[4,8,16] Rather, he claimed it had no particular (probabilistic) interpretation per se, and was only an index of evidence in the same sense that, say, a foot is an index of length or a pound is an index of weight. So when reporting results, one should simply note the magnitude of the *P*-value. No particular interpretation is necessary (just as we do not "interpret" a measurement of someone's height).

The first clinical illustration is an example of reporting findings by way of a significance test. The researchers

reported the *P*-value to communicate the strength of the evidence in the data and characterized the evidence as marginal. Additional recommendations for modifying clinical practice based on these data would be discussed later and would have to be weighted in light of possible side effects and costs.

Once reported, researchers are supposed to ruminate on the magnitude of the *P*-value and consider the scientific context. Fisher suggested that the 5% level (p <0.05) could be used as a scientific benchmark for concluding that fairly strong evidence exists against the null hypothesis.[16] However he never intended this level to be an absolute threshold. The strength of evidence does not jump from one category to the next (weak to moderate to strong, etc.). Rather, it is on a continuum that gradually and smoothly increases in the same way that measurements of height and weight do. Whereas a *P*-value of 0.049 represents fairly strong evidence, so to does a one of 0.055 or 0.07, albeit to a lesser degree. Fisher also maintained that scientific context was critical. A *P*-value of 0.05, for example, might lead to the recommendation that additional experiments be performed in one circumstance, whereas that same value could be taken as ironclad evidence of an effect in another situation.

Fisher referred to the process of drawing conclusions from data as inductive inference.[18] His significance tests were to be used to measure and summarize the strength of statistical evidence in the data against a particular hypothesis. Unfortunately, significance tests are seldom distinguished from hypothesis tests, leading to confusion and inaccuracies in reporting and interpreting *P*-values.

Hypothesis Testing, Types I and II Errors, and Making Decisions

The hypothesis test is an altogether different animal from the significance test. In 1933, J. Neyman and E. Pearson introduced hypothesis testing as an alternative to significance testing.[19] Their idea was to take a mathematical rule for choosing between two hypotheses, say a null and alternative hypothesis, and determine how often this rule would lead researchers astray. They were then able to find the one rule that was optimal in the sense that it led researchers astray least often.

A hypothesis test is formulated in terms of two hypotheses and two error rates. Unlike significance testing, an alternative hypothesis must be specified. If the null hypothesis is true but the test tells us to choose the alternative, we have made an error of the first kind (type I); if the alternative is true and the test tells us to choose the null hypothesis, we have made an error of the second kind (type II). Types I and II error *rates* are the probability of making these errors. Neyman and Pearson were able to identify the one decision rule that minimized the type II error rate when all of the rules under consideration had the same type I error rate.[19] Therefore researchers need only specify the type I error rate, after which they could then construct the Neyman–Pearson rejection rule to tell them what hypothesis to reject, with the knowledge that this rule would naturally keep the type II error rate as small as possible. Note that the type I error

rate is also represented by the tail area probability under the null hypothesis.

A hypothesis test is classically conducted in two stages. The first stage occurs before data are collected and the second stage occurs after data are collected. In the first stage, null and alternative hypotheses are specified, the type I error rate is chosen, and Neyman–Pearson's rejection rule is determined. After collecting data, one checks to see if the null hypothesis is rejected. One then reports only whether or not the null hypothesis is rejected, and what type I error level was used (important: the magnitude of the P-value is irrelevant here; all that matters is the particular action taken). Because this procedure is concerned only with taking appropriate action, it is a model of inductive behavior, not inductive inference.

The second clinical research illustration is an example of how hypothesis tests are reported. Only rejection of the null hypothesis and the 5% significance level (type I error rate) are reported. The concept of statistical evidence is absent here. For error rates to have their intended meanings, investigators (and all clinicians) should accept the null hypothesis as false and modify their behavior accordingly. There is no ruminating over the results here! Of course, this does not happen because the reporting of scientific results is not about making decisions, but about collecting, summarizing, and reevaluating evidence. However, the ability to design experiments that controlled certain types of errors was so attractive that researchers and scientists found another way to take advantage of this framework. What they did is use the hypothesis-testing framework to design a study, and the significance-testing framework to report and interpret study results.

The problem with this approach is that interpretation of error rates is meaningless because researchers no longer act in direct accordance with the hypothesis test. If researchers and scientists always acted in accordance with the test (they always accepted the hypothesis that was not rejected), they would be led to only accept incorrect hypotheses as often as types I and II error rates specified. But once a hypothesis test is conducted, researchers are not supposed to reevaluate those hypotheses. When they do, true error rates become inflated and uncontrollable. There are ways to avoid this, but they require determining how many times the hypothesis under consideration will be tested. Once that is known, types I and II errors can be adjusted proactively (Goodman provides a nice overview of this issue).[20] Unfortunately this is often impossible to determine in practice.

Confusion Reigns as Significance and Hypothesis Testing Are Married

Fisher was the first to see the writing on the wall. He understood that these two testing procedures, so similar in mechanics and terminology yet so drastically different in purpose, could easily be confused as one. Neyman and Pearson also understood what was at stake, and all three wrote extensively about how their procedures were differ-

ent. But much of what they said went unheeded because of acrimony between Neyman and Fisher.

In retrospect, it was probably only a matter of time before these two frameworks were merged. Why? Because each one has a key concept that the other lacks. Significance testing provides a measure of the strength of evidence, but does not address how often that measure may be misleading. Hypothesis testing does just the opposite: error rates provide a sense of how often one may be mislead, but they do not represent or measure the strength of evidence in the data. So it seems only natural that researchers would want to use both frameworks to design good experiments and characterize evidence in the data.

The end result of this marriage is an ad hoc methodology based loosely on partial definitions, choice principles, and key quantities from both testing procedures. But because of this, this methodology gives rise to irresolvable controversies such as those surrounding adjustment for multiple comparisons and multiple looks at accumulating data.[4,20] Moreover, the union of these two testing frameworks invites misinterpretation of key quantities, such as interpreting the P-value as a post hoc type I error rate.[4,8]

To Adjust or Not to Adjust

One example of the confusion that this merger has created concerns adjustments of the tail area probability for multiple comparisons or multiple looks at accumulating data.[4,20] Should error rates or P-values be adjusted for multiple comparisons or repeated testing of accumulating data? Following the line of reasoning laid down here, the answer is that error rates should be adjusted, but P-values should not (as long as they are not given a probabilistic interpretation).

Adjustment of error rates for multiple comparisons or repeated testing is necessary to keep that error rate controlled. Otherwise it inflates with each examination of the data (with each opportunity to make an error). But P-values are different; they measure the strength of evidence in the data, which should depend only on the data at hand and not on how many other examinations were conducted or are planned. The key idea is this: repeated testing increases the propensity for some results along the way to be misleading, so we attempt to control that propensity by adjusting the error rate. But repeated testing does not change how we interpret what the data themselves represent in terms of statistical evidence. Why should it? If two experiments happen to collect exactly the same data, they should have exactly the same amount of statistical evidence. But the propensity for each set of data to be misleading could be different depending on the experimental design.

We should note that this is a hotly debated topic. Statisticians sometimes point out that in certain situations P-values can be very misleading if they are not adjusted for repeated testing,[21,22] while others note that adjusting P-values can lead to statistical conclusions that violate common sense (two researchers with exactly the same data can then end up with different P-values and therefore different

assessments of the strength of evidence in the data).[4,23] Again, following our line of reasoning, *P*-values are an index of the strength of evidence in the data and should therefore not be adjusted.

P-Values, Sample Size, and Statistical versus Clinical Significance

Implicit in the significance testing framework is the concept that equal *P*-values from two different experiments represent the same strength of evidence against the null hypothesis, regardless of sample sizes. Fisher strongly believed this,[24] and many others supported this idea.[25] This concept was named the α postulate.[23]

But the α postulate is wrong. A given *P*-value does not have a fixed meaning independent of sample size. When two studies have equal *P*-values, some experts held that evidence from the one with the smaller sample was actually stronger,[26] but others came to the exact opposite conclusion.[27] In fact, both interpretations are correct, depending on if the exact *P*-value is reported (p = 0.003) or if it is reported only to be less than some fixed threshold (p <0.01).[28] Such discrepancies are certainly not reassuring, and they indicate that *P*-values may not be the best measure of statistical evidence available. Alternative viewpoints are published elsewhere.[2,7,29]

The take-home message is that statistical significance depends critically on sample size, and clinical significance should always be considered. What is absent from our two illustrations is exactly this: a definition of what differences are clinically important for the comparison. The first example shows 2-fold improvement, but the difference is not statistically significant. The second example has lots of statistical power due to large sample, and a statistically significant difference, but the difference is not likely to be clinically significant. Although the *P*-value provides a measure of statistical significance, it fails to connect that measure directly to the estimated magnitude of the effect. That is why many experts prefer the use of effect measures (relative risk, odds ratio, etc.) and CIs instead.[9–13] In this sense, *P*-values tell only half the story.

Many experts argued that researchers put too much emphasis on *P*-values because researchers tend to focus on achieving statistical significance with limited regard to clinical significance. Instead, one should focus on estimation and CIs. The *P*-value has limited value in terms of assessing the magnitude of an effect because it depends heavily on sample size. Hence the common warning that statistical significance does not imply clinical significance. For this reason, leaders in the field have suggested that *P*-values be replaced, or at the least augmented with effect estimates and CIs.

How has this been translated into practice in medical journals? As a result of the marriage of significance and hypothesis testing, journals and researchers often divide results into two categories: statistically significant (positive) studies and results that are not statistically significant (negative). This has resulted in two common and potentially seri-

ous consequences: clinically significant differences noted in small studies are considered nonsignificant and are ignored, and all statistically significant findings are assumed to result from real treatment effects and are assumed to be important (but may be clinically insignificant).[30,31] It is very important to consider sample size, and in this sense CIs can be extremely helpful. We will return to this, after first clearing up some misinterpretations of *P*-values.

Common Misinterpretations of *P*-Values

P-values are routinely misused and misinterpreted in clinical research. In fact, they are easy to misinterpret because their seemingly natural or desired interpretation is often technically incorrect. As discussed earlier, they do not require a specific interpretation other than as index of the strength of evidence in the data. However, they are often confused with type I errors, resulting in some of the following mistaken interpretations:

1. The *P*-value is the likelihood that findings are due to chance.
2. The *P*-value is 0.06; therefore, there is a 94% chance that a real difference exists.
3. With a low *P*-value (p <0.001), the findings must be true.
4. The lower the *P*-value, the stronger the evidence for an effect.
5. Equal *P*-values represent the same amount of evidence against the null hypothesis.
6. The *P*-value is the probability that the null hypothesis is true given the data.

Interpretations 1, 2, and 6 attempt to interpret the *P*-value as a probability. As we noted earlier, if the *P*-value is an index of the evidence, it should not be interpreted in this way. Interpretation 6 is wrong because the *P*-value is calculated assuming that the null hypothesis is true, so it cannot represent the uncertainly about the null hypothesis as well.[32] Interpretations 4 and 5 both rest on the validity of the α postulate. These statements are true only when sample sizes are the same among experiments. Otherwise they may or may not be true. Finally, interpretation 3 is also not correct. Here context is very important, and even with very low *P*-values we will never be absolutely sure that the null hypothesis is false.

Effect Sizes and Confidence Intervals

As mentioned, many experts and journal editors prefer effect sizes and CIs to *P*-values. The *P*-value is often taken to split findings into two groups, statistically significant (p <0.05) and insignificant (p >0.05) because of confusion with hypothesis testing. This makes little sense from an evidential perspective and promotes superficial thinking about research findings.[33] For example, does it makes scientific sense to adopt a new therapy because the *P*-value of a single study was 0.049, and at the same time ignore the results of another therapy because the *P*-value was 0.051? Certainly not, although that is exactly what one does under

the hypothesis-testing framework. Effect sizes and confidence intervals can provide an indirect assessment of the strength of evidence in the data. However, they do not suffer from problems with sample size, as *P*-values do, and for that reason alone are a welcome improvement.

As Grimes succinctly stated, "the fundamental problem with the *P*-value is that it conveys no information about the size of differences or associations found in studies."[9] Reporting an effect size, such as a relative risk (RR) or odds ratio, provides a measure of the strength of the association. A relative risk of 4.0 implies a 4-fold increase, and a relative risk of 30 is a 30-fold increase. Consider our clinical examples: in the first example we had a 2-fold difference (RR = 2.0) that was statistically insignificant. In the second, the difference was a modest 10% improvement (RR = 1.1), but was statistically significant. Surely the reader will agree that an RR of 2.0 is more clinically meaningful than an RR of 1.1. But how good is this estimate? To answer this question, we use a CI.

The effect measure, or point estimate (RR), is the best estimate for the true but unknown effect under investigation. A CI for the effect (usually at the 95% level) indicates how good that estimate is by providing a range of uncertainty in the point estimate. The width of the interval indicates precision in our data; wide CIs are less precise and narrow ones more precise. Thus we have an indication of clinical significance because the interval will include only clinical important effects, only clinically unimportant effects, or both types of effects. The first and second cases are self-explanatory; these intervals clearly indicate what the data say. But the third case is not so clear, and it indicates that more data should be collected. The problem with looking only at *P*-values is that a small figure could be associated with any one of these three examples depending on the sample size and null hypothesis being tested.

The connection between *P*-values and CIs is as follows. If a 95% CI includes the null effect, the *P*-value is greater than 0.05 and a test of the null hypothesis would fail to reject. (The null effect is that specified under the null hypothesis. In our example the null effect would be that the relative risk is unity.) If a 95% CI excludes the null effect (RR = 1.0 in our examples) the associated *P*-value for testing that null hypothesis will be less than 0.05 and a test of the null hypothesis will reject. In fact, a 95% CI can be thought of as a collection of all null hypotheses that would not reject at the 5% level. It is most easily interpreted as a collection of hypotheses that are best supported by, or consistent with, the data at the 95% level.

In our first illustration, the relative risk was 2.0 and the 95% CI was 0.98 to 4.1. From this we know that a hypothesis test would not reject at the 5% level and that the *P*-value would be greater than 0.05 (we know this because the CI includes 1.0). Here the CI includes values that are not clinically important. In the second illustration, where the null hypothesis was rejected and the result deemed statistically significant, the RR was only 1.1 and the 95% CI was 1.01 to 1.2. This CI indicates that if there is any effect at all, it is likely to be a small improvement. Clearly, presenting a CI greatly improves the dissemination and characterization of results.

Should I Measure the Evidence or Make a Decision?

The natural tendency, we believe, is for scientists and researchers to want to measure and summarize evidence. Hypothesis testing puts too much emphasis on a single statistically significant study, without regard to costs and benefits of the therapy under consideration. Moreover, published medical research reports often provide no firm evidence for decision making. Each report contributes incrementally to an existing body of knowledge. Increasing recognition of this fact is reflected in the growth of formal methods of research synthesis,[8] including presentation of an updated meta-analysis in the discussion section of original research reports.[9] When presented in this manner, prior evidence is based on results of reports addressing the same issue, and new data are added to the body of evidence. All forms of evidence (animal studies, different epidemiologic study designs, etc.) should be considered as one weighs the evidence, and this is not done in the hypothesis-testing framework. The discussion section of a research report should put the results in the context of other evidence in the medical literature to arrive at a logical conclusion.

In our opinion, conclusions and recommendations for decision making should not be based on results of a hypothesis test alone. Evidence for clinical practice should be based on all available evidence, the strength of the association, the precision of this estimate, potential public health benefits (and harms), and economic considerations. The *P*-value plays an important role in this respect and that is why it will just not disappear, despite it deficiencies. Science needs a way to measure and summarize the strength of the evidence in data objectively. Without a better option, the *P*-value is here to stay; however, we can avoid some of the problems associated with it by presenting effect sizes and CIs for the effect under investigation.

Summary

The arbitrary dichotomies of research findings as statistically significant and insignificant, which result from the pure hypothesis-testing approach, are often not helpful scientifically and can lead to problems. Researchers want to measure and summarize the strength of evidence in the data. This is what *P*-values are used for, not for making decisions. And although they may not be the best measure available, they are the standard of statistical care at this time. However researchers, experts, and journal editors are increasingly encouraging the use of effect measures (point estimates) and CI estimation over both hypothesis and significance testing. Effect sizes and CIs often provide critical information that should not be overlooked and can help to reduce the dependence on *P*-values in cases where they are likely to be problematic. New research findings must be put into the context of existing knowledge. Medical evidence for decision making should rely on a synthesis of

existing research studies, and the contribution of the new data presented to the current evidence in the literature.

References

1. Goodman SN: Toward evidence-based medical statistics. 1. The probability value fallacy. Ann Intern Med 130:995–1004, 1999

2. Goodman SN: Toward evidence-based medical statistics. 2. The Bayes factor. Ann Intern Med 130:1005–1013, 1999

3. Berger J: Could Fisher, Jeffreys and Neyman have agreed on testing? Stat Sci 18:1–32, 2003

4. Royall RM: Statistical Evidence: A Likelihood Paradigm. London, Chapman & Hall, 1997

5. Hubbard R, Bayarri MJ: Confusions over measures of evidence (p's) versus errors (alphas's) in classical statistical testing. Am Statistician 57:171–182, 2003

6. Goodman SN, Royall RM: Evidence and scientific research. Am J Public Health 78:1568–1574, 1988

7. Goodman SN: Of p-values and Bayes: A modest proposal. Epidemiology 12:295–297, 2001

8. Goodman SN: P-values, hypothesis tests, and likelihood: Implications for epidemiology of a neglected historical debate. Am J Epidemiol 137:485–496, 1993

9. Grimes DA: The case for confidence intervals. Obstet Gynecol 80:865–866, 1992

10. Altman DG: Confidence intervals in research evaluation. ACP Journal Club 116(suppl 2):A28–9, 1992

11. Berry G: Statistical significance and confidence intervals [editorial]. Med. J Aust 144:618–619, 1986

12. Braitman LE: Confidence intervals extract clinically useful information from data [editorial]. Ann Intern Med 108:296–298, 1988

13. Simon R: Confidence interval for reporting the results of clinical trials. Ann Intern Med 105:429–435, 1986

14. Fisher RA: Statistical Methods for Research Workers. Edinburgh, Oliver & Boyd, 1925

15. Fisher RA: The Design of Experiments. Edinburgh, Oliver & Boyd, 1935

16. Fisher RA: Statistical Methods for Research Workers, 13th ed. New York, Hafner, 1958

17. Fisher RA: Statistical Methods and Scientific Inference, 2nd ed. New York, Hafner, 1959

18. Fisher RA: The logic of inductive inference. J R Stat Soc Series B 98:39–54, 1935

19. Neyman J, Pearson E: On the problem of the most efficient tests of statistical hypotheses. Phil Trans R Soc A 231:289–337, 1933

20. Goodman SN: Multiple comparisons, explained. Am J Epidemiol 147:807–812, 1998

21. Armitage P: Some developments in the theory and practice of sequential medical trials. Proceedings of the 5th Berkeley symposium. Math Stat Probl 4:791–804, 1967

22. Armitage P, McPherson CK, Rowe BC: Repeated significance tests on accumulating data. J R Stat Soc 132:235–244, 1969

23. Cornfield J: Sequential trials, sequential analysis, and the likelihood principle. Am Statistician 29:18–23, 1966

24. Fisher RA: Statistical Methods for Research Workers, 5th ed. London, Oliver & Boyd, 1934

25. Bergson J: Tests of significance considered as evidence. J Am Stat Assoc 37:325–335, 1942

26. Lindley DV, Swott WF: New Cambridge Elementary Statistical Tables. London, Cambridge University Press, 1984

27. Peto R, Pike MC, Armitage P, et al: Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design. Br Med J 34:585–612, 1976

28. Royall RM: The effect of sample size on the meaning of significant tests. Am Statistician 40:313–315, 1986

29. Blume JD: Likelihood methods for measuring statistical evidence. Stat Med 21:2563–2599, 2002

30. Sterne JAC, Smith GD: Sifting the evidence—What's wrong with significance tests? Lancet 322:226–231, 2001

31. Freiman JA, Chalmers TC, Smith H, et al: The importance of beta, the type II error, and sample size in the design and interpretation of the randomized controlled trial. N Engl J Med 299:690–694, 1978

32. Cohen J: The earth is round (p<0.05). Am Psychol 49:997–1003, 1994

33. Poole C: Beyond the confidence interval. Am J Public Health 77:195–199, 1987