ELSEVIER

Seminars in
HEMATOLOGY

A Dirty Dozen: Twelve *P*-Value Misconceptions

Steven Goodman

The *P* value is a measure of statistical evidence that appears in virtually all medical research papers. Its interpretation is made extraordinarily difficult because it is not part of any formal system of statistical inference. As a result, the *P* value's inferential meaning is widely and often wildly misconstrued, a fact that has been pointed out in innumerable papers and books appearing since at least the 1940s. This commentary reviews a dozen of these common misinterpretations and explains why each is wrong. It also reviews the possible consequences of these improper understandings or representations of its meaning. Finally, it contrasts the *P* value with its Bayesian counterpart, the Bayes' factor, which has virtually all of the desirable properties of an evidential measure that the *P* value lacks, most notably interpretability. The most serious consequence of this array of *P*-value misconceptions is the false belief that the probability of a conclusion being in error can be calculated from the data in a single experiment without reference to external evidence or the plausibility of the underlying mechanism.
Semin Hematol 45:135-140

The *P* value is probably the most ubiquitous and at the same time, misunderstood, misinterpreted, and occasionally miscalculated index[1,2] in all of biomedical research. In a recent survey of medical residents published in *JAMA*, 88% expressed fair to complete confidence in interpreting *P* values, yet only 62% of these could answer an elementary *P*-value interpretation question correctly.[3] However, it is not just those statistics that testify to the difficulty in interpreting *P* values. In an exquisite irony, none of the answers offered for the *P*-value question was correct, as is explained later in this chapter.

Writing about *P* values seems barely to make a dent in the mountain of misconceptions; articles have appeared in the biomedical literature for at least 70 years[4-15] warning researchers of the interpretive *P*-value minefield, yet these lessons appear to be either unread, ignored, not believed, or forgotten as each new wave of researchers is introduced to the brave new technical lexicon of medical research.

It is not the fault of researchers that the *P* value is difficult to interpret correctly. The man who introduced it as a formal research tool, the statistician and geneticist R.A. Fisher, could not explain exactly its inferential meaning. He proposed a rather informal system that could be *used*, but he never could describe straightforwardly what it meant from an inferential standpoint. In Fisher's system, the *P* value was to be used as a rough numerical guide of the strength of evidence against the null hypothesis. There was no mention of "error rates" or hypothesis "rejection"; it was meant to be an evidential tool, to be used flexibly within the context of a given problem.[16]

Fisher proposed the use of the term "significant" to be attached to small *P* values, and the choice of that particular word was quite deliberate. The meaning he intended was quite close to that word's common language interpretation—something worthy of notice. In his enormously influential 1926 text, *Statistical Methods for Research Workers,* the first modern statistical handbook that guided generations of biomedical investigators, he said:

> *Personally, the writer prefers to set a low standard of significance at the 5 percent point . . . . A scientific fact should be regarded as experimentally established only if a properly designed experiment rarely fails to give this level of significance.*[17]

In other words, the operational meaning of a *P* value less than .05 was merely that one should *repeat the experiment*. If subsequent studies also yielded significant *P* values, one could conclude that the observed effects were unlikely to be the result of chance alone. So "significance" is merely that: worthy of attention in the form of meriting more experimentation, but not proof in itself.

The *P* value story, as nuanced as it was at its outset, got incomparably more complicated with the introduction of the machinery of "hypothesis testing," the mainstay of current practice. Hypothesis testing involves a null and alternative hypothesis, "accepting and rejecting" hypotheses, type I and

Departments of Oncology, Epidemiology, and Biostatistics, Johns Hopkins Schools of Medicine and Public Health, Baltimore, MD.
Address correspondence to Steven Goodman, MD, MHS, PhD, 550 N Broadway, Suite 1103, Baltimore, MD, 21205. E-mail: Sgoodman@jhmi.edu

0037-1963/08/$-see front matter
doi:10.1053/j.seminhematol.2008.04.003

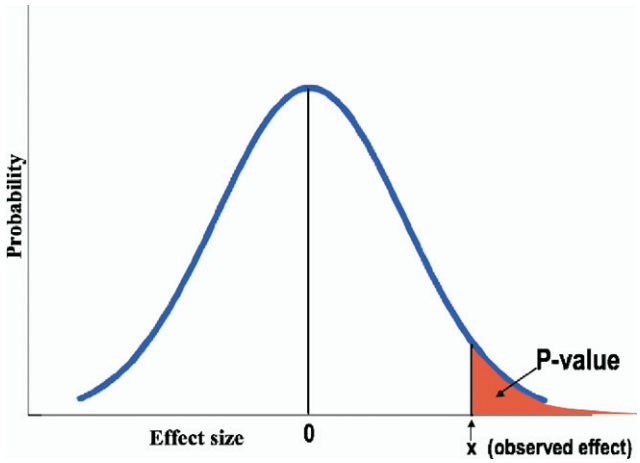

Figure 1 Graphical depiction of the definition of a (one-sided) *P* value. The curve represents the probability of every observed outcome under the null hypothesis. The *P* value is the probability of the observed outcome (x) plus all "more extreme" outcomes, represented by the shaded "tail area."

II "error rates," "power," and other related ideas. Even though we use *P* values in the context of this testing system today, it is not a comfortable marriage, and many of the misconceptions we will review flow from that unnatural union. In-depth explanation of the incoherence of this system, and the confusion that flows from its use can be found in the literature.[16,18-20] Here we will focus on misconceptions about how the *P* value should be interpreted.

The definition of the *P* value is as follows—in words: *The probability of the observed result, plus more extreme results, if the null hypothesis were true;* in algebraic notation: Prob(X $\geq x$ | Ho), where "X" is a random variable corresponding to some way of summarizing data (such as a mean or proportion), and "*x*" is the observed value of that summary in the current data. This is shown graphically in Figure 1.

We have now mathematically defined this thing we call a *P* value, but the scientific question is, what does it *mean*? This is not the same as asking what people *do* when they observe $P \leq .05$. That is a custom, best described sociologically. Actions should be motivated or justified by some conception of foundational meaning, which is what we will explore here.

Because the *P* value is not part of any formal calculus of inference, its meaning is elusive. Below are listed the most common misinterpretations of the *P* value, with a brief discussion of why they are incorrect. Some of the misconceptions listed are equivalent, although not often recognized as such. We will then look at the *P* value through a Bayesian lens to get a better understanding of what it means from an inferential standpoint.

For simplicity, we will assume that the *P* value arises from a two-group randomized experiment, in which the effect of an intervention is measured as a difference in some average characteristic, like a cure rate. We will not explore the many other reasons a study or statistical analysis can be misleading, from the presence of hidden bias to the use of improper models; we will focus exclusively on the *P* value itself, under ideal circumstances. The null hypothesis will be defined as the hypothesis that there is no effect of the intervention (Table 1).

Misconception #1: *If P=.05, the null hypothesis has only a 5% chance of being true.* This is, without a doubt, the most pervasive and pernicious of the many misconceptions about the *P* value. It perpetuates the false idea that the data alone can tell us how likely we are to be right or wrong in our conclusions. The simplest way to see that this is false is to note that the *P* value is calculated under the assumption that the null hypothesis is true. It therefore cannot simultaneously be a probability that the null hypothesis is false. Let us suppose we flip a penny four times and observe four heads, two-sided $P = .125$. This does not mean that the probability of the coin being fair is only 12.5%. The only way we can calculate that probability is by Bayes' theorem, to be discussed later and in other chapters in this issue of *Seminars in Hematology*.[21-24]

Misconception #2: *A nonsignificant difference (eg, P >.05) means there is no difference between groups.* A nonsignificant difference merely means that a null effect is statistically consistent with the observed results, together with the range of effects included in the confidence interval. It does not make the null effect the most likely. The effect best supported by the data from a given experiment is always the observed effect, regardless of its significance.

Misconception #3: *A statistically significant finding is clini-*

Table 1 Twelve *P*-Value Misconceptions

1	*If P = .05, the null hypothesis has only a 5% chance of being true.*
2	*A nonsignificant difference (eg, P ≥.05) means there is no difference between groups.*
3	*A statistically significant finding is clinically important.*
4	*Studies with P values on opposite sides of .05 are conflicting.*
5	*Studies with the same P value provide the same evidence against the null hypothesis.*
6	*P = .05 means that we have observed data that would occur only 5% of the time under the null hypothesis.*
7	*P = .05 and P ≤.05 mean the same thing.*
8	*P values are properly written as inequalities (eg, "P ≤.02" when P = .015)*
9	*P = .05 means that if you reject the null hypothesis, the probability of a type I error is only 5%.*
10	*With a P = .05 threshold for significance, the chance of a type I error will be 5%.*
11	*You should use a one-sided P value when you don't care about a result in one direction, or a difference in that direction is impossible.*
12	*A scientific conclusion or treatment policy should be based on whether or not the P value is significant.*

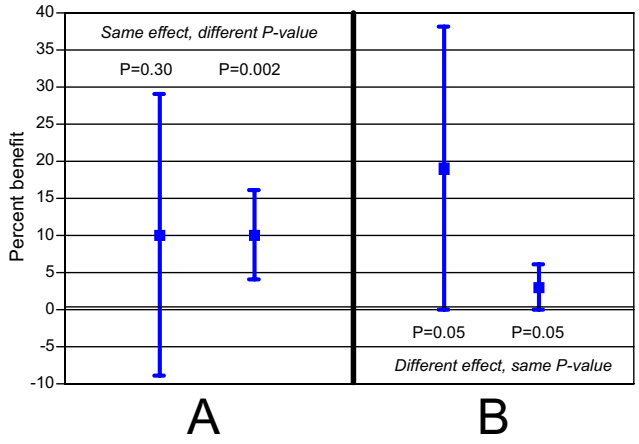

Figure 2 Figure showing how the *P* values of very different significance can arise from trials showing the identical effect with different precision (A, Misconception #4), or how same *P* value can be derived from profoundly different results (B, Misconception #5).

cally important. This is often untrue. First, the difference may be too small to be clinically important. The *P* value carries no information about the magnitude of an effect, which is captured by the effect estimate and confidence interval. Second, the end point may itself not be clinically important, as can occur with some surrogate outcomes: response rates versus survival, CD4 counts versus clinical disease, change in a measurement scale versus improved functioning, and so on.[25-27]

Misconception #4: *Studies with P values on opposite sides of .05 are conflicting.* Studies can have differing degrees of significance even when the estimates of treatment benefit are *identical*, by changing only the precision of the estimate, typically through the sample size (Figure 2A). Studies statistically conflict only when the difference between their results is unlikely to have occurred by chance, corresponding to when their confidence intervals show little or no overlap, formally assessed with a test of heterogeneity.

Misconception #5: *Studies with the same P value provide the same evidence against the null hypothesis.* Dramatically different observed effects can have the same *P* value. Figure 2B shows the results of two trials, one with a treatment effect of 3% (confidence interval [CI], 0% to 6%), and the other with an effect of 19% (CI, 0% to 38%). These both have a *P* value of .05, but the fact that these mean different things is easily demonstrated. If we felt that a 10% benefit was necessary to offset the adverse effects of this therapy, we might well adopt a therapy on the basis of the study showing the large effect and strongly reject that therapy based on the study showing the small effect, which rules out a 10% benefit. It is of course also possible to have the same *P* value even if the lower CI is not close to zero.

This seeming incongruity occurs because the *P* value defines "evidence" relative to only one hypothesis—the null. There is no notion of positive evidence—if data with a $P = .05$ are evidence against the null, what are they evidence *for*? In this example, the strongest evidence for a benefit is for 3% in one study and 19% in the other. If we quantified evidence in a relative way, and asked which experiment provided greater evidence for a 10% or higher effect (versus the null), we would find that the evidence was far greater in the trial showing a 19% benefit.[13,18,28]

Misconception #6: *$P = .05$ means that we have observed data that would occur only 5% of the time under the null hypothesis.* That this is not the case is seen immediately from the *P* value's definition, the probability of the observed data, *plus more extreme data*, under the null hypothesis. The result with the *P* value of exactly .05 (or any other value) is the most probable of all the other possible results included in the "tail area" that defines the *P* value. The probability of any individual result is actually quite small, and Fisher said he threw in the rest of the tail area "as an approximation." As we will see later in this chapter, the inclusion of these rarer outcomes poses serious logical and quantitative problems for the *P* value, and using comparative rather than single probabilities to measure evidence eliminates the need to include outcomes other than what was observed.

This is the error made in the published survey of medical residents cited in the Introduction,[3] where the following four answers were offered as possible interpretations of $P > .05$:

a. The chances are greater than 1 in 20 that a difference would be found again if the study were repeated.
b. The probability is less than 1 in 20 that a difference this large could occur by chance alone.
c. The probability is greater than 1 in 20 that a difference this large could occur by chance alone.
d. The chance is 95% that the study is correct.

The correct answer was identified as "c", whereas the actual correct answer should have read, "The probability is greater than 1 in 20 that a difference this large *or larger* could occur by chance alone."

These "more extreme" values included in the *P*-value definition actually introduce an operational difficulty in calculating *P* values, as more extreme data are by definition *unobserved* data. What "could" have been observed depends on what experiment we imagine repeating. This means that two experiments with identical data on identical patients could generate different *P* values if the imagined "long run" were different. This can occur when one study uses a stopping rule, and the other does not, or if one employs multiple comparisons and the other does not.[29,30]

Misconception #7: *$P = .05$ and $P \leq .05$ mean the same thing.* This misconception shows how diabolically difficult it is to either explain or understand *P* values. There is a big difference between these results in terms of weight of evidence, but because the same number (5%) is associated with each, that difference is literally impossible to communicate. It can be calculated and seen clearly only using a Bayesian evidence metric.[16]

Misconception #8: *P values are properly written as inequalities (eg, "$P \leq .02$" when $P = .015$).* Expressing all *P* values as inequalities is a confusion that comes from the combination of hypothesis tests and *P* values. In a hypothesis test, a pre-set "rejection" threshold is established. It is typically set at $P = .05$, corresponding to a type I error rate (or "alpha") of 5%. In such a test, the only relevant information is whether the

difference observed fell into the rejection region or not, for example, whether or not $P \leq .05$. In that case, expressing the result as an inequality ($P \leq .05$ *v* $P > .05$) makes sense. But we are usually interested in how *much* evidence there is against the null hypothesis; that is the reason *P* values are used. For that purpose, it matters whether the *P* value equals .50, .06, .04 or .00001. To convey the strength of evidence, the exact *P* value should always be reported. If an inequality is used to indicate merely whether the null hypothesis should be rejected or not, that can be done only with a pre-specified threshold, like .05. *The threshold cannot depend on the observed P value*, meaning we cannot report "$P < .01$" if we observe $P = .008$ and the threshold was .05. No matter how low the *P* value, we must report "$P < .05$." But rejection is very rarely the issue of sole interest. Many medical journals require that very small *P* values (eg, $< .001$) be reported as inequalities as a stylistic issue. This is ordinarily not a big problem except in situations where literally thousands of statistical tests have been done (as in genomic experiments) when many very small *P* values can be generated by chance, and the distinction between the small and the extremely small *P* values are important for proper conclusions.

Misconception #9: *P = .05 means that if you reject the null hypothesis, the probability of a type I error is only 5%.* Now we are getting into logical quicksand. This statement is equivalent to Misconception #1, although that can be hard to see immediately. A type I error is a "false positive," a conclusion that there is a difference when no difference exists. If such a conclusion represents an error, then by definition there is no difference. So a 5% chance of a false rejection is equivalent to saying that there is a 5% chance that the null hypothesis is true, which is Misconception #1.

Another way to see that this is incorrect is to imagine that we are examining a series of experiments on a therapy we are certain is effective, such as insulin for diabetes. If we reject the null hypothesis, the probability that rejection is false (a type 1 error) is zero. Since all rejections of the null hypothesis are true, it does not matter what the *P* value is. Conversely, if we were testing a worthless therapy, say copper bracelets for diabetes, all rejections would be false, regardless of the *P* value. So the chance that a rejection is right or wrong clearly depends on more than just the *P* value. Using the Bayesian lexicon, it depends also on our a priori certitude (or the strength of external evidence), which is quantified as the "prior probability" of a hypothesis.

Misconception #10: *With a P = .05 threshold for significance, the chance of a type I error will be 5%.* What is different about this statement from Misconception #9 is that here we are looking at the chance of a type I error *before* the experiment is done, not after rejection. However, as in the previous case, the chance of a type I error depends on the prior probability that the null hypothesis is true. If it is true, then the chance of a false rejection is indeed 5%. If we know the null hypothesis is false, there is no chance of a type I error. If we are unsure, the chance of a false positive lies between zero and 5%.

The point above assumes no issues with multiplicity or study design. However, in this new age of genomic medicine, it is often the case that literally thousands of implicit hypotheses can be addressed in a single analysis, as in comparing the expression of 5,000 genes between diseased and non-diseased subjects. If we define "type I error" as the probability that any of thousands of possible predictors will be falsely declared as "real," then the *P* value on any particular predictor has little connection with the type I error related to the whole experiment. Here, the problem is not just with the *P* value itself but with the disconnection between the *P* value calculated for one predictor and a hypothesis encompassing many possible predictors. Another way to frame the issue is that the search through thousands of predictors implies a very low prior probability for any one of them, making the posterior probability for a single comparison extremely low even with a low *P* value. Since the 1 − (posterior probability) is the probability of making an error when declaring that relationship "real," a quite low *P* value still carries with it a high probability of false rejection.[31,32]

Misconception #11: *You should use a one-sided P value when you don't care about a result in one direction, or a difference in that direction is impossible.* This is a surprisingly subtle and complex issue that has received a fair amount of technical discussion, and there are reasonable grounds for disagreement.[33-38] But the operational effect of using a one-sided *P* value is to increase the apparent strength of evidence for a result based on considerations not found in the data. Thus, use of a one-sided *P* value means the *P* value will incorporate attitudes, beliefs or preferences of the experimenter into the assessment of the strength of evidence. If we are interested in the *P* value as a measure of the strength of evidence, this does not make sense. If we are interested in the probabilities of making type I or type II errors, then considerations of one-sided or two-sided rejection regions could make sense, but there is no need to use *P* values in that context.

Misconception #12: *A scientific conclusion or treatment policy should be based on whether or not the P value is significant.* This misconception encompasses all of the others. It is equivalent to saying that the magnitude of effect is not relevant, that only evidence relevant to a scientific conclusion is in the experiment at hand, and that both beliefs and actions flow directly from the statistical results. The evidence from a given study needs to be combined with that from prior work to generate a conclusion. In some instances, a scientifically defensible conclusion might be that the null hypothesis is still probably true even after a significant result, and in other instances, a nonsignificant *P* value might still lead to a conclusion that a treatment works. This can be done formally only through Bayesian approaches. To justify actions, we must incorporate the seriousness of errors flowing from the actions together with the chance that the conclusions are wrong.

These misconceptions do not exhaust the range of misstatements about statistical measures, inference or even the *P* value, but most of those not listed are derivative from the 12 described above. It is perhaps useful to understand how to measure true evidential meaning, and look at the *P* value from that perspective. There exists only one calculus for quantitative inference—Bayes' theorem—explicated in more

depth elsewhere and in other articles in this issue. Bayes' theorem can be written in words in this way:

Odds of the null hypothesis after obtaining the data

= Odds of the null hypothesis before obtaining the data

× Bayes' factor

or to use more technical terms:

$$\text{Posterior odds } (H_0, \text{ given the data}) = \text{Posterior odds } (H_0, \text{ given the data}) \times \frac{\text{Prob(Data, under } H_0)}{\text{Prob(Data, under } H_A)}$$

where Odds = probability/(1 − probability), H_0 = null hypothesis, and H_A = alternative hypothesis.

It is illuminating that the *P* value does not appear anywhere in this equation. Instead, we have something called the Bayes' factor (also called the likelihood ratio in some settings), which is basically the same as the likelihood ratio used in diagnostic testing.[24,39] It measures how strongly the observed data are predicted by two competing hypotheses, and is a measure of evidence that has most of the properties that we normally mistakenly ascribe to the *P* value. Table 2 summarizes desirable properties of an evidential measure, and contrasts the likelihood ratio to the *P* value. The main point here is that our intuition about what constitutes a good measure of evidence is correct; what is problematic is that the *P* value has few of them. Interested readers are referred to more comprehensive treatments of this contrast, which show, among other things, that the *P* value greatly overstates the evidence against the null hypothesis.[40] (See article by Sander Greenland in this issue for more complete discussion of Bayesian approaches[41]). Table 3 shows how *P* values can be compared to the strongest Bayes' factors that can be mustered for that degree of deviation from the null hypothesis. What this table shows is that (1) *P* values overstate the evidence against the null hypothesis, and (2) the chance that rejection of the null hypothesis is mistaken is far higher than is generally appreciated even when the prior probability is 50%.

One of many reasons that *P* values persist is that they are part of the vocabulary of research; whatever they do or do not mean, the scientific community feels they understand the rules with regard to their use, and are collectively not familiar enough with alternative methodologies or metrics. This was discovered by the editor of the journal *Epidemiology* who tried to ban their use but was forced to abandon the effort after several years.[42]

In the meantime, what is an enlightened and well-meaning researcher to do? The most important foundational issue to appreciate is that there is no number generated by standard methods that tells us the probability that a given conclusion is right or wrong. The determinants of the truth of a knowledge claim lie in combination of evidence both within and outside a given experiment, including the plausibility and evidential support of the proposed underlying mechanism. If that mechanism is unlikely, as with homeopathy or perhaps intercessory prayer, a low *P* value is not going to make a treatment based on that mechanism plausible. It is a very rare single experiment that establishes proof. That recognition alone prevents many of the worst uses and abuses of the *P* value. The second principle is that the size of an effect matters, and that the entire confidence interval should be considered as an experiment's result, more so than the *P* value or even the effect estimate. The confidence interval incorporates both the size and imprecision in effect estimated by the data.

There hopefully will come a time when Bayesian measures of evidence, or at least Bayesian modes of thinking, will supplant the current ones, but until then we can still use standard measures sensibly if we understand how to reinterpret them through a Bayesian filter, and appreciate that our inferences must rest on many more pillars of support than the study at hand.

Table 2 Evidential Properties of Bayes' Factor Versus *P* Value

Evidential Property	*P* Value	Bayes' Factor
Information about effect size?	No	Yes
Uses only observed data?	No	Yes
Explicit alternative hypothesis?	No	Yes
Positive evidence?	No	Yes
Sensitivity to stopping rules?	Yes	No
Easily combined across experiments?	No	Yes
Part of formal system of inference?	No	Yes

Table 3 Correspondence Between *P* Value, Smallest Bayes' Factor, and Posterior Probability of an "Even Odds" Hypothesis

P Value	Smallest Bayes' Factor	Smallest Posterior Probability of H_0 When Prior Probability = 50%
.10	.26	21%
.05	.15	13%
.03	.10	9%
.01	.04	4%
.001	.005	.5%

References

1. Garcia-Berthou E, Alcaraz C: Incongruence between test statistics and P values in medical papers. BMC Med Res Methodol 4:13, 2004
2. Andersen B: Methodological Errors in Medical Research. Oxford, UK, Blackwell Science, 1990
3. Windish DM, Huot SJ, Green ML: Medicine residents' understanding of the biostatistics and results in the medical literature. JAMA 298:1010-1022, 2007
4. Berkson J: Tests of significance considered as evidence. J Am Stat Assoc 37:325-35, 1942
5. Mainland D: The significance of "nonsignificance." Clin Pharm Ther 5:580-586, 1963
6. Mainland D: Statistical ritual in clinical journals: Is there a cure? —I. Br Med J 288:841-843, 1984
7. Edwards W, Lindman H, Savage LJ: Bayesian statistical inference for psychological research. Psych Rev 70:193-242, 1963

8. Diamond GA, Forrester JS: Clinical trials and statistical verdicts: Probable grounds for appeal. Ann Intern Med 98:385-394, 1983
9. Feinstein AR: P-values and confidence intervals: Two sides of the same unsatisfactory coin. J Clin Epidemiol 51:355-360, 1998
10. Feinstein AR: Clinical biostatistics. XXXIV. The other side of 'statistical significance': Alpha, beta, delta, and the calculation of sample size. Clin Pharmacol Ther 18:491-505, 1975
11. Rothman K: Significance questing. Ann Intern Med 105:445-447, 1986
12. Pharoah P: How not to interpret a P value? J Natl Cancer Inst 99:332-333, 2007
13. Goodman SN, Royall R: Evidence and scientific research. Am J Public Health 78:1568-1574, 1988
14. Braitman L: Confidence intervals extract clinically useful information from data. Ann Intern Med 108:296-298, 1988
15. Goodman SN: Towards evidence-based medical statistics, I: The P-value fallacy. Ann Intern Med 130:995-1004, 1999
16. Goodman SN: P-values, hypothesis tests and likelihood: Implications for epidemiology of a neglected historical debate. Am J Epidemiol 137: 485-496, 1993
17. Fisher RA: Statistical Methods for Research Workers. Oxofrd, UK, Oxford University Press, 1958
18. Royall R: Statistical Evidence: A Likelihood Paradigm. London, UK, Chapman & Hall, 1997
19. Gigerenzer G, Swijtink Z, Porter T, Daston L, Beatty J, Kruger L: The Empire of Chance. Cambridge, UK, Cambridge University Press, 1989
20. Lehmann EL: The Fisher, Neyman-Pearson theories of testing hypotheses: One theory or two? J Am Stat Assoc 88:1242-1249, 1993
21. Lilford RJ, Braunholtz D: For debate: The statistical basis of public policy: A paradigm shift is overdue. BMJ 313:603-607, 1996
22. Greenland S: Bayesian perspectives for epidemiological research: I. Foundations and basic methods. Int J Epidemiol 35:765-775, 2006
23. Greenland S: Randomization, statistics, and causal inference. Epidemiology 1:421-429, 1990
24. Goodman SN: Towards evidence-based medical statistics, II: The Bayes' factor. Ann Intern Med 130:1005-1013, 1999
25. Rothman KJ: A show of confidence. N Engl J Med 299:1362-1363, 1978
26. Gardner MJ, Altman DG: Confidence intervals rather than p values: Estimation rather than hypothesis testing. Stat Med 292:746-750, 1986
27. Simon R: Confidence intervals for reporting results of clinical trials. Ann Intern Med 105:429-435, 1986
28. Goodman SN: Introduction to Bayesian methods I: Measuring the strength of evidence. Clin Trials 2:282-290, 2005
29. Berry DA: Interim analyses in clinical trials: Classical vs. Bayesian approaches. Stat Med 4:521-526, 1985
30. Berger JO, Berry DA: Statistical analysis and the illusion of objectivity. Am Sci 76:159-165, 1988
31. Ioannidis JP: Why most published research findings are false. PLoS Med 2:e124, 2005
32. Ioannidis JP: Genetic associations: False or true? Trends Mol Med 9:135-138, 2003
33. Goodman SN: One or two-sided P-values? Control Clin Trials 9:387-388, 1988
34. Bland J, Altman D: One and two sided tests of significance. BMJ 309: 248, 1994
35. Boissel JP: Some thoughts on two-tailed tests (and two-sided designs). Control Clin Trials 9:385-386, 1988 (letter)
36. Peace KE: Some thoughts on one-tailed tests. Biometrics 44:911-912, 1988 (letter)
37. Fleiss JL: One-tailed versus two-tailed tests: Rebuttal. Control Clin Trials 10:227-228, 1989 (letter)
38. Knottnerus JA, Bouter LM: The ethics of sample size: Two-sided testing and one-sided thinking. J Clin Epidemiol 54:109-110, 2001
39. Kass RE, Raftery AE: Bayes' factors. J Am Stat Assoc 90:773-795, 1995
40. Berger JO, Sellke T: Testing a point null hypothesis: The irreconcilability of P-values and evidence. J Am Stat Assoc 82:112-122, 1987
41. Greenland S: Bayesian interpretation and analysis of research results. Semin Hematol (this issue)
42. Lang JM, Rothman KJ, Cann CI: That confounded P-value. Epidemiology 9:7-8, 1998