Point of view

A farewell to P-values?

A recent paper in the medical literature by Sterne and Davey-Smith[1] has focused attention on some of the problems of interpretation of significance/hypothesis tests; in particular, the meaning and status of P-values associated with these tests. What were these concerns: that the division of results into "significant" and "non-significant" according to a P-value = 0.05 was arbitrary and not in accordance with the prescriptions of the founders of statistical inference; that the P-value is misinterpreted as the probability that the null hypothesis is true; that, as the absolute value of the P-value indexes the level of evidence against the null hypothesis, measures of effect should attract a P-value of 0.001, in preference to 0.05, where the evidence against the null hypothesis is not strong; Bayesian approaches to reporting of results may have advantage; and "significance" should not be a primary claim of the reporting of results, which should be accompanied by (90%) confidence intervals and interpreted in the context of the type of study and other available evidence.

As the authors observe, these matters are not new and have repeatedly surfaced in the literature of various scientific disciplines since the establishment of the "testing" paradigm in the 1920s and 1930s by RA Fisher and J Neyman & E Pearson.[2] Two volumes, separated by almost 30 years and authored from within the behavioural science disciplines: "The significance test controversy - a reader"[3] (published in 1969) and "What if there were no significance tests"[4] (published in 1997), further attest to these controversies. In a provocatively entitled paper, "Two cheers for P-values", Stephen Senn, in a "limited defence of P-values", noted that "P-values are a practical success, but a critical failure. Scientists the world over use them, but scarcely a statistician can be found to defend them. Bayesians in particular find them ridiculous….".[5] In 1996, Nester suggested that "statisticians would be unwise to seek the limelight in any forthcoming 75th anniversary, centennial or tricentennial celebrations of hypothesis testing".[6] Rindskopf asked why "Given the many attacks on it, null hypothesis testing…(was not) …dead",[7] and the demise of the P-value has been rhetorically reported.[8]

Within the medical literature similar sentiments have been expressed, as reflected in the titles of certain lead articles: "Confidence intervals rather than P values: estimation rather than hypothesis testing",[9] "That confounded P-value"[10] and "Are all significant P values created equal?".[11] That this was not merely an academic question, was revealed by a decision of the editor of the journal Epidemiology:[12] "When writing for Epidemilogy, you can also enhance your prospects if you omit tests of statistical significance….we do not publish them at all".[13] In psychology and the social sciences, the tone of discourse (against P-values) has at times been shrill, as noted by Nickerson in a recent exhaustive review.[14] In the biological[15] and econometric literature,[16] others have added to the chorus of complaint. How did this come about or is it all "a tempest in a tea pot"?[14]

History: the paradigm established

From our current medical perspective, "testing", P-values and Type I and II errors appear non-problematic; a "single, unified, uncontroversial means of statistical inference",[17] and the history of the development of statistical inference in the 20th century, a remote echo of current concerns.[18] The state of statistics in the first decade of the 20th century has been described as "an unexplored archeological site"[19] and the construction of the "testing" paradigm, first by Fisher and then Neyman-Pearson, was self-consciously defined with respect to the 19th century Bayesian dominance of "inverse probability". It was paradoxical that Gosset, the inventor of the *t*-test (1909), which initiated hypothesis testing,[20] was a Bayesian.[21]

It is obviously difficult to relive the impact of the Fisherian revolution upon the statistical practice of the first decades of the 20th century, but we may be assured that it was fundamental, although the full significance of the first edition (1925) of the classic "Statistical methods for Research Workers" was , perhaps not surprisingly, "not immediately recognized".[22] Fisher and Gosset, in fact, cooperated in calculating tables for the *t*-distribution presented (along with χ^2 and *z*-transformation) in the book.[16] The initial "common currency" of significant at 5% and 1% may well have been related to the fact that Fisher's tables (copied subsequently to many text books) were given for P-values of 0.01 and 0.05, partly in deference to the copyright limitations of the journal Biometrika, edited by Karl Pearson, the founder of the χ^2 test.[23]

R.A. Fisher's position as the "founder" of modern statistics is presumably secure;[24] the comments of Savage (a Bayesian) attest to this: "It would be more economical to list the few statistical topics in which he displayed no interest than those in which he did".[25] The Fisherian *significance* test, deriving from inductive inference, established a null hypothesis (H_0) and used discrepancies in the data to reject the null hypothesis: that is, H_0 posited a sample coming from a hypothetical (infinite) population with a known sampling distribution and H_0 was rejected if the sample estimate deviated from the mean of the sampling distribution by more than a specified criterion (the level of significance;

formally, Pr (x|H_0)).[17] The P-value then, was the (tail area) probability of obtaining a result equal to or more extreme than what was actually observed;[21] for, say a P value of 0.05, it was *not* that the null hypothesis had a probability of (only) 5%. Under an assumption that the null hypothesis was true, it could *not* then be assumed that the P value was a "direct measure of the probability that the null hypothesis is false".[26] A P value of $\leq$ 0.05 on the null hypothesis indicated, according to Fisher, that: "Either an exceptionally rare chance has occurred or the theory is not true". It did not imply that the investigator accepted being deceived one in twenty occasions, rather it suggested what should be ignored: "all experiments in which significant results are not obtained". Fisher's further advice, oft quoted and ignored, was that "If one in twenty does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty…or one in a hundred".[17] A subtle, but important point, was that the inference from the P value involved only one hypothesis and was partially based on unobserved data in the tail region (of the sampling distribution); thus the "likelihood" of a hypothesis, deriving from the data , was not, at the same time, the "probability of being true".[18,27] That is, P-values were not to be misinterpreted as posterior probabilities, a Bayesian proposition.[28]

The statistical methodology of Neyman & Pearson, articulated primarily in two papers in 1928 and 1933,[29] sought to revise and improve upon Fisher's formulation, but from a deductive position, a paradigmatic difference, which from the Fisherian viewpoint of "rigorous inferences from the particular to the general", was a function of a certain mathematical "bias" of Neyman. Although somewhat distrustful of mathematicians ("Statistical methods for Research Workers" contained no formal mathematical proofs or lemmas), Fisher was in fact a Cambridge trained mathematician.[25] The Fisher-Neyman rivalry (Pearson later "distanced" himself from the Neyman-Pearson paradigm, more so when Neyman relocated to Berkley) was somewhat of a *cause-celebre*,[30] although the influence of Fisher's "Statistical methods for Research Workers" on the Neyman-Pearson enterprise was acknowledged by the latter,[20] specifically the tabulation of the three distributions mentioned above. Neyman, for his part, charged Fisher with a persistent inability to operate with concepts;[31] Fisher's circumlocutions were also a cause of irritation to those sympathetic to his view-point, such as Kempthorne: "The last sentence, particularly, leads me to the view that Fisher was talking on a plane barely understandable to the rest of humanity".[32]

It was Neyman-Pearson methodology which formulated the now familiar two competing hypotheses paradigm, the null (H_0) and the alternate (H_A). This involved the probability of committing two kinds of errors with respect to the null hypothesis; false rejection (Type I or α error) and false acceptance (Type II or β error). Power, defined as (1-β) or the P(rejecting H_0 | a particular alternative hypothesis), was introduced as a new and critical concept. Within the Fisherian schema, there was a notion of "sensitivity" in detecting departures from the null, but no formal concept of power;[33] although Barnard has argued otherwise.[34] The α error was, in theory, prescribed *prior* to data collection and the focus was on minimizing β errors, subject to the bound upon α. What in effect was established were rules for making decisions between two hypotheses ("inductive behavior", although this behavioral aspect may only have been heuristic or hypothetical),[35] on the basis that in the "long run of experience, we shall not be too often wrong".[17] Thus for Neyman-Pearson, there was a tension between the control of long term error rates and judgment of the status of the individual experiment.[20] The α and β error rates defined a rejection *region* for a test statistic; the significance level, α, was the "probability of a set of future outcomes", represented by the "tail" area of the null distribution. In principle, Neyman-Pearson theory also avoided an arbitrary element in Fisher's approach, the decision regarding the test statistic.[5] The Neyman-Pearson fundamental lemma guaranteed the existence of an optimal (uniformly) Most Powerful-α test;[36] for a simple H_0 tested against a simple alternative H_A, the optimal test criterion was the likelihood ratio (LR) test.[23,37] The question of Fisher's approach to the alternative hypothesis is one of some difficulty: arguments that a small probability $p(E|h_o)$ of event E is "not enough *per se* to discredit the null hypothesis h_0" have been "forcefully" advanced.[33] In particular, Berkson's oft cited paper from 1941 in which the question was posed: "If an event has occurred, the definitive question is not, "Is this an event which would be rare if H_0 is true?" but "Is there an alternative hypothesis under which the event would be relatively frequent?" If there is no plausible alternative at all, the rarity is quite irrelevant to a decision…".[38] Undoubtedly Fisher considered the 'alternative' as obligatory, as revealed in conversation with Kruskal and Savage in the 1950's, where the former recalls that "..Fisher agreed that, yes, naturally one had to think about distributions for the sample other than that of the hypothesis under test and why were we making such a fuss about an elementary and trivial question".[39]

That significance testing could provide statistical inference, rather than behavioural decisions, was the gulf that separated Fisher from (early) Neyman-Pearson. Some modification of the Neyman-Pearson position on this did occur, both internally (Pearson's description of a statistical test as a "means of learning"[40] and Neyman's subsequent equivocations on the matter of inference[41]) and externally, such as the position of Lehmann, in the classic 1959 volume "Testing statistical hypotheses", that in a hypothesis test the "information will be used for guidance…In such cases the emphasis is on the inference…".[42]

A data-based P-value is a random variable with a distribution, under the null hypothesis and for continuous test statistics, uniform over the interval [0,1], regardless of the size of the study.[43,44] Under the alternative hypothesis, the distribution of the P-value is skewed and is a function of both sample size ("a natural concept of power")[46] and the distribution of the test statistic that is used. P-values are therefore not α-error rates,[5,17,20,26] although both are tail-area probabilities under the null hypothesis. As Berger and Delampady note: "*P* values are not a repetitive error rate, at least in any real sense. A Neyman-Pearson error probability, α, has the actual frequentist interpretation that a long series of α level tests will reject no more than 100α% of true H_0, but the data dependent *P* values have no such interpretation".[46]

For a fixed, pre-specified α, the Neyman-Pearson decision rule "could be defined equivalently in terms of the P-value"[17] but what would be of interest was the fact that $P < \alpha$, not the specific value of P. It is ironic that standard applied statistical practice lies easily with an amalgam of the two methods, although from an alternate perspective, this amalgam may be considered as one of statistics "greater triumphs".[47] Tests of significance, as reported in journals, would appear to follow Neyman-Pearson "formally" but Fisher "philosophically" and practically.[33]

In the debate about the utility of P-values versus confidence intervals (CI), it is often forgotten that CI were introduced by Neyman in 1937,[48] and were considered by Neyman, and commentators, as integral to the overall theory of hypothesis testing,[49] which embodied the frequentist theory of repeated sampling (an anathema to Fisher[50]). Thus for a 95% CI of a parameter θ, the interpretation is that in an infinite number of repetitions of a study, an exact proportion (95%) of all such intervals would enclose θ. Once the data has been collected and a single 95% CI has been calculated, the probability that θ lies within this CI is now 0 or 1. That is, a 95% CI is not equivalent to a 95% probability interval (which has a Bayesian interpretation).[51-54]

Besides the theory of testing, the second great divide between Neyman and Fisher was that of confidence versus fiducial intervals, the latter being based on fiducial probability,[55] introduced by Fisher as an alternative to Bayesian posterior probabilities.[55] Fiducial probability had, as its basis, certain sufficient statistics (*F* and *t* statistics and correlations) which contained all the information in a sample relevant to population parameters (inference from sample to population). Although agreement can be demonstrated between CI and fiducial intervals; the classic paper establishing so called exact (or Clopper-Pearson) CI of the binomial was entitled "The use of confidence or fiducial limits illustrated in the case of the binomial";[56] the latter has not stood the test of Neyman's withering attacks,[31,57,58] nor time (Fisher's "biggest blunder"[19]). It is of interest, however, to record that Fisher (as early as 1935) apparently "recognized that 'confidence intervals' are 'only another way of saying that, by a certain test of significance, some kinds of hypothetical possibilities are to be rejected, while others are not' ".[59]

History: the paradigm revised

The literature in response to the Fisher and Neyman-Pearson paradigm is, not surprisingly, enormous in breadth and detail, both from within the statistical,[35,60-64] and applied scientific disciplines. One of the more systematic and useful developments is that associated with DR Cox, in which P-values are treated as "rough tools for inspecting data".[65] The perspective, developed fully in the 1974 volume of Cox and Hinkley,[66] is one of eclecticism, the central theme being that "it is fruitful to contemplate problems formulated in different depths of detail and to use different methods accordingly…the most primitive formulation is that for a pure significance test, where only the null hypothesis under test need be explicitly formulated, and the richest formulation is that for Bayesian decision analysis…".[67] Null hypotheses are not viewed as undifferentiated species, rather, are divided into plausible (close to the truth) and dividing (divide the range of possibilities into qualitatively different types) hypotheses, which may also be further sub-divided and specified. Statistical strategies, the use of significance tests and the actual P-value level, are determined contingent upon these hypothesis specifications.[68,69]

A significance test (measuring the consistency of the data with a null hypothesis) has the following form: a function $t = t(y)$ of the observations exists, such that, the larger $t(y)$, the greater the inconsistency of y (the observed vector of responses) with H_0. $T = t(y)$ is the test statistic (a random variable). If the distribution of T is known (when H_0 is true), then the level of significance $p_{\text{obs}} = \text{pr}\ (T \geq t_{\text{obs}}\text{: } H_0)$. The result of such a test is a significance level (not a decision); p_{obs} is a "guide, and no more, to interpretation" (albeit the mathematical connection between p_{obs} and "critical regions of pre-assigned size", that is, Neyman-Pearson testing).[66,68,70] A similar approach was also recommended by Kempthorne.[71] Of interest, Cox, in a somewhat sympathetic response to the paper by Sterne and Davey-Smith,[1] suggested that "To distinguish several types of hypotheses that might be tested helps to understand the issues".[72]

The problem revisited

Where then do we stand? In a response to the debate over the paper by Sterne and Davey-Smith,[1] Berger,[73] outlined potential problems with hypothesis testing and it is useful to consider some of these:

1. *P-values are misunderstood*. The frequent misrepresentation of P-values and hypothesis testing, especially in textbooks, has been repeatedly documented.[74,75] Such, as with other statistical misrepresentations, is not an argument for their abolition.
2. *P-values as a measure of support*. Sterne and Davey-Smith,[1] suggest a graded level of evidence against the null hypothesis, indexed by the P-value; such scales date back to the 1970s.[70] A corollary to this is the so called α-postulate of Cornfield (rejected by him in favour of likelihood ratios), that "All hypotheses rejected at the same critical level have equal amounts of evidence against them".[76] However, the question of sample size for equal P-values needs to be considered; a number of commentators have argued that for, say a P-value of 0.05, there is stronger evidence against H_0 for a small sample than a large one.[27,38,70,77-79] Schervish also demonstrated that "the interpretation of a particular value on the scale of support, such as the popular .05, must vary with the hypothesis" and was "unable to construct a consistent interpretation of the P-value as anything similar to a measure of support for its hypothesis".[44]
3. *P-values are associated with rigid cut-off values*. A flexible (eclectic) attitude to tests and P-values, associated with the approach of Cox and Kempthorne, has been outlined above. As opposed to Berger, it would indeed appear reasonable that there "should be no sharp distinction made between cases having a P-value of say 4.9% and those having a P-value of 5.1%- a distinction forced by the language of confidence interval testing".[80] In the presence of a bewildering array of possible statistics, the advice of Kempthorne to "Look at it"[80] seems apposite; similar to the admonitions of the adherents of the likelihood principle[81] and the Bayesians.
4. *P-values are the wrong measure of evidence*. From the Bayesian perspective, P-values overstate the evidence against the null hypothesis and other methods to adduce evidence (likelihood ratios) may be of more utility.[82]

   In a frequently cited paper by J.O. Berger and Sellke, it was shown (two sided testing a normal mean) that with a P-value of 0.05, the posterior probability of the null was at least 0.30 for any objective prior distribution.[83] However, in the one-sided setting, where the different geometry of H_0 and H_A was not operative, the discrepancy, Bayesian posterior probability versus P-value, was no longer evident.[84] Technically, this 'geometry' relates to the fixing of a (prior) probability mass on the null and varying it on the alternative; Casella and R. L. Berger suggest that the discrepancy between P-values and $P(H_0|x)$ is a function of the large (50%) prior probability mass placed on H_0 by J.O. Berger and co-workers,[46,83] and conclude that "there is agreement between P-values and Bayesian interval null calculations in the more typical situation in which small prior probability is assigned to H_0".[85] As Casella and R. L. Berger note, "We would be surprised if most researchers would place even a 10% prior probabilty on H_0 ", in accord with the sentiments of Meehl, who maintained that the point null hypothesis is "…[quasi-] always false in biological and social science".[86] P-values and posterior probabilities are not necessarily in competition and any difference in conclusions reached does not "….by itself invalidate either measure".[74] Of interest, J.O. Berger and co-workers recently proposed calibrating P-values such that they may be interpreted in either a Bayesian fashion ($B(p) = -e.p \log(p)$, when $p < 1/e$) or a frequentist way ($\alpha(p) = (1+[-e.p \log(p)]^{-1})^{-1}$.[87]

Goodman, again from a Bayesian perspective, calculated the replication probability (using an uninformative prior) of trials at a P-value of 0.05; this was found to be 50% and lower than "expected" (by non-statisticians).[88] Senn has subjected the import of this finding to close scrutiny.[5,21] Firstly, from a Bayesian perspective, P-values are not unreasonable given an uninformative prior. However, the problem is that "…the 'uninformative' prior is rarely appropriate…it is not possible to survive as a Bayesian on uninformative priors…".[21] Secondly, the requirement that a single significant P-value should entail near certainty that a second will follow, is deemed by Senn to be an undesirable property: "Anticipated evidence is not evidence, nor do we want it to be. To expect that it is, is to make exactly the same mistake that physicians make in saying, 'the result was not significant, p = 0.09, because the trial was too small'".[5] This being said of the general problem of replicability, empirical studies have suggested that the P-value does provide "a continuous measure that has an orderly and monotonic mapping onto confidence in the replicabiltiy of a null hypothesis rejection"[89] and statistically significant exact replication (SSER) may be a useful interpretative measure.[90]

P-values and CI

There would appear to be considerable virtue in reporting both P-values and CI, on the basis that singular statements such as P < 0.05, or P = nil sig, convey little useful information, although for a 100(1-α)% CI, it must be remembered that any violation of the assumptions that effect the true vale of α (obviously) effect CI precision.[91] From the Bayesian perspective, Lindley has summarized the position thus: "significance tests, as inference procedures, are better replaced by estimation methods…it is better to quote a confidence or credible interval….estimation procedures provide more information…..Nevertheless there remain cases where significance test have an advantage…".[92] Numerous papers within the medical literature have attested to the utility of CI;[93-95] in the epidemiological literature, the P-value has been condemned as confounded, in that the information "mixed" in the P-value should be separately reported: the size of the effect (estimated by, say, the risk ratio) and the precision of the estimate (described by the SE or CI).[10] Poole has suggested that for epidemiological measures such as relative risk, the estimates least influenced by chance are those with narrow confidence intervals, not low P-values.[96] However, as pointed out by Feinstein, P-values and CI methods are essentially reciprocal and do not provide "an evaluation of substantive importance for the 'big' or 'small' magnitude of the observed distinction…they offer no guidance for the basic quantitative scientific appraisals that depend on purely descriptive rather than inferential boundaries..".[97] Similarly, Poole suggested that CI "are usually taken as nothing more than tests of significance" and proposed that the complete P-value (that is, the graph of all possible P-values or CI) or likelihood function be used for the "main result of an epidemiological study".[98]

An interesting case study of the interpretation of "CI without P-values" and a focal point for a lively exchange in the American Journal of Public Health in the mid 1980's on the role of P-values, was the response of Fleiss to a relatively small case-control study[99] in which all 12 reported CI for summary odds ratios included 1 (extending from 0.4 to 17); yet the associations were variously described as 'strong' 'negative' and 'positive'. Fleiss asked the not unreasonable question "There is no gainsaying that tests of significance have been abused, but at least they have the virtue of providing explicit, pre-specifiable criteria for inferring that an association is real. This is not the case with confidence intervals, at least as far as the paper in question is concerned….I would appreciate learning just what criteria were employed to conclude that the ….associations were 'strong', 'negative' and 'positive'".[100] The authors acknowledged the small study size, and suggested that the reader should be "cautious in generalizing our results", but even when not "statistically significant" point estimates may "significantly add to our understanding".[101] The latter position would appear to distance itself considerably from the Fisherian requirement of "rigorous uncertainty".[30]

The question of the "propriety' of associations and/or claims of efficacy also resonates with the reporting of drug trials;[102] the case for confidence intervals as estimates of effect has been well argued[103] and indeed, would appear to be non-controversial. In the presence of competing claims and professional enthusiasts, one can, with V.W. Berger, question the likelihood of being mislead[73] and find P-values eminently applicable to control this probability,[104] not withstanding the utility of other (for example, Bayesian) approaches.[105]

Overview

Efron summarized the major reasons why, as opposed to the Bayesian 19th century, Fisherian and Neyman-Pearson ideas have held sway in the 20th: ease of use, model building, division of labour (parts of a complicated problem may be addressed separately) and objectivity.[106] Thus, despite the belief that P-values are dead and buried (by some journals), we would agree with Fleiss that significance tests are "alive and well".[107]

Post-script

As this is published in an Australian journal, we provide the following further information: Ronald Aylmer Fisher was born February 17 1890 in East Finchley, London and died July 29, 1962 in Adelaide, South Australia. His ashes are interred in St Peter's Cathedral in the latter city.

J. L. MORAN
Department of Intensive Care Medicine, Queen Elizabeth Hospital, Woodville, SOUTH AUSTRALIA

P. J. SOLOMON
School of Mathematical Sciences, University of Adelaide, Adelaide, SOUTH AUSTRALIA

REFERENCES

1. Sterne JA, Davey SG. Sifting the evidence-what's wrong with significance tests? BMJ 2001;322:226-231.
2. Gigerenzer G, Swijtink Z, Porter T, Daston L, Beatty J, Kruger L. The Empire of Chance : How Probability Changed Science and Everyday Life. New York: Cambridge University Press; 1989.
3. Morrison DE, Henkel RE. The significance test controversy-a reader. Chicago,Ill: Aldine Publishing; 1969.
4. Harlow LL, Mulaik SA, Steiger JH. What if there were no significance tests. Hillsdale, NJ: Lawrence Erlbaum Associates; 1997.
5. Senn S. Two cheers for P-values? J Epidemiol Biostat 2001;6:193-204.

6. Nester MR. An applied Statistician's creed. Applied Statistics 1996;45:401-410.
7. Rindskopf DM. Testing "small", not null, hypotheses: Classical and Bayesian approaches. In: Harlow LL, Muliak SA, Steiger JH. What if there were no significance tests? Hillsdale NJ: Lawrence Erlbaum Associates; 1997: 319-322.
8. Evans SJ, Mills P, Dawson J. The end of the p value?. Br Heart J 1988;60:177-180.
9. Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. Br Med J (Clin Res Ed) 1986;292:746-750.
10. Lang JM, Rothman KJ, Cann CI. That confounded P-value. Epidemiology 1998;9:7-8.
11. Browner WS, Newman TB. Are all significant P values created equal? The analogy between diagnostic tests and clinical research. JAMA 1987;257:2459-2463.
12. Epidemiology. http://www epidem com/pt/re/epidemiology/home htm 2004.
13. Rothman KJ. Writing for epidemiology. Epidemiology 1998;9:333-337.
14. Nickerson RS. Null hypothesis significance testing: a review of an old and continuing controversy. Psychol Methods 2000;5:241-301.
15. Johnson DH. The insignificance of statistical significance testing. Journal of Wildlife Management 1999;63:763-772.
16. Keuzenkamp HA, Magnus JR. On tests and significance in econometrics. Journal of Econometrics 1995;67:5-24.
17. Hubbard R, Bayarri, MJ. Confusion over measures of evidence (p's) versus errors (a's) in classical statistical testing. The American Statistician 2003;57:171-182.
18. Goodman SN. p values, hypothesis tests, and likelihood: implications for epidemiology of a neglected historical debate. Am J Epidemiol 1993;137:485-496.
19. Efron B. R.A. Fisher in the 21st Century. Statistical Science 1998;13:95-122.
20. Lehmann EL. The Fisher, Neyman-Pearson theories of testing hypotheses: One theory or two. Journal of the American Statistical Association 1993;88:1242-1249.
21. Senn S. A comment on replication, p-values and evidence, S.N.Goodman, Statistics in Medicine 1992; 11:875-879. Stat Med 2002;21:2437-2444.
22. Yates F. The influence of Statistical Methods for Research Workers on the development of the science of statistics. Journal of the American Statistical Association 1951;46:19-34.
23. Barnard GA. Must clinical trials be large? The interpretation of P-values and the combination of test results. Stat Med 1990;9:601-614.
24. Rao CR. R.A. Fisher: The founder of modern statistics. Statistical Science 1992;7:34-48.
25. Savage LJ. On rereading RA Fisher. The Annals of Statistics 1976;4:441-500.
26. Goodman SN. Toward evidence-based medical statistics. 1: The P value fallacy. Ann Intern Med 1999;130:995-1004.
27. Gibbons JD, Pratt JW. P-values: Interpretation and Methodology. The American Statistician 1975;29:20-25.
28. De Groot MH. Doing what comes naturally: interpreting a tail area as a posterior probability or as a likelihood ratio. Journal of the American Statistical Association 1973;68:966-969.
29. Neyman J, Pearson ES. Joint Statistical Papers. London, UK: Cambridge University Press; 1967.
30. Marks HM. Rigorous uncertainty: Why RA Fisher is important. Int J Epidemiol 2003;32:932-937.
31. Neyman J. Silver Jubilee of my dispute with Fisher. Journal of the Operational Research Society of Japan 1961;3:145-154.
32. Kempthorne O. Some aspects of experimental inference. Journal of the American Statistical Association 1966;61:11-34.
33. Johnstone DJ. Tests of significance in theory and practice. The Statistician 1986;35:491-504.
34. Barnard GA. Discussion of 'Tests of significance in theory and practice' by D.J. Johnstone. The Statistician 1986;35:499-502.
35. Birnbaum A. The Neyman-Pearson theory as decision theory, and as inference theory; with a criticism of the Lindley-Savage argument for Bayesian theory. Synthese 1977;36:19-49.
36. Pena EA, Rohatgi VK. Most powerful test. In: Armitage P, Colton T. Encyclopedia of Biostatistics. New York: John Wiley & Sons, Inc; 1998: 2703-2706.
37. Neyman J, Pearson ES. On the problem of the most efficient test of statistical hypotheses. Philosophical Transactions of the Royal Society of London Series A: Mathematical and Physical Sciences 1933;231:289-337.
38. Berkson J. Tests of significance considered as evidence. Journal of the American Statistical Association 1947;37:325-335.
39. Kruskal W. The significance of Fisher: a review of [Box JF] R.A. Fisher: the life of a scientist. Journal of the American Statistical Association 1980;75:1019-1030.
40. Pearson ES. Statistical concepts and their relation to reality. Journal of the Royal Statistical Society, Series B 1955;17:204-207.
41. Johnstone DJ. On the interpretation of hypothesis tests followign Neyman and Pearson. In: Viertl R. Probability and Bayesian Statistics. New York, NY: Plenum Press; 1987: 267-277.
42. Lehman EL. Testing statistical hypotheses. New York: John Wiley and Sons; 1959.
43. Hung HM, O'Neill RT, Bauer P, Kohne K. The behavior of the P-value when the alternative hypothesis is true. Biometrics 1997;53:11-22.
44. Schervish MJ. P values: what they are and what they are not. The American Statistician 1996;50:203-206.
45. Schweder T. A significance version of the basic Neyman-Peason theory for scientific hypothesis testing. Scandinavian Journal of Statistics 1988;15:225-242.
46. Berger JO, Delampady M. Testing precise hypotheses. Statistical Science 1987;2:317-352.
47. Carlton MA. Discussion to "Confusion over measures of evidence (p's) versus errors (a's) in classical statistical testing". The American Statistician 2003;57:179-181.
48. Neyman J. Outline of a theory of statistical estimation based on the classical theory of probabilty. Philosophical Transactions of the Royal Society of London Series A: Mathematical and Physical Sciences 1937;236:333-380.
49. Neyman J. Frequentist probability and frequentist statistics. Synthese 1977;36:97-131.

50. Fisher RA. Statistical methods and scientific induction. Journal of the Royal Statistical Society, Series B 1955;17:69-78.
51. Goodman SN. Confidence limits vs power calculations. Epidemiology 1994;5:266-268.
52. Kupper LL. Estimation, Interval. In: Armitage P, Colton T. Encyclopedia of Biostatistics. New York: John Wiley & Sons, Inc; 1998: 1391-1394.
53. Macdonald RR. The Incompleteness of Probability Models and the Resultant Implications for Theories of Statistical Inference. Understanding Statistics 2002;1:167-189.
54. Young KD, Lewis RJ. What is confidence? Part 2: Detailed definition and determination of confidence intervals. Ann Emerg Med 1997;30:311-318.
55. Seidenfeld T. Fiducial probability. In: Armitage P, Colton T. Encyclopedia of Biostatistics. New York: John Wiley & Sons, Inc; 1998: 1510-1515.
56. Clopper CJ, Pearson, E S. The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika 1934;26:404-413.
57. Neyman J. Fiducial argument and the theory of confidence intervals. Biometrika 1941;32:128-150.
58. Neyman J. Note on an article by Sir Ronald Fisher. Journal of the Royal Statistical Society, Series B 1956;18:288-294.
59. Dempster AP. Comment on RA Fisher in the 21st century by B Efron. Statistical Science 1998;13:120-121.
60. Berger JO. Could Fisher, Jeffreys and Neyman have agreed on Testing. Statistical Science 2003;18:1-32.
61. Hacking I. The logic of statistical inference. Cambridge: Cambridge University Press; 1965.
62. Kyburg HE Jr. The logical foundations of Statistical Inference. Boston, USA: D. Reidel publishing Company; 1974.
63. Spielman S. A refutation of the Neyman-Pearson theory of testing. British Journal of Philosophy of Science 1973;24:201-222.
64. Spielman S. The logic of tests of significance. Philosophy of Science 1974;41:211-226.
65. Salsburg D. Hypothesis testing. In: Armitage P, Colton T. Encyclopedia of Biostatistics. New York: John Wiley & Sons, Inc; 1998: 1969-1976.
66. Cox DR, Hinkley DV. Theoretical Statistics. London: Chapman & Hall; 1974.
67. Cox DR. Foundations of statistical inference: The case for eclecticism. Australian Journal of Statistics 1978;20:43-59.
68. Cox DR. The role of significance tests. Scandanavian Journal of Statistics 1977;4:49-70.
69. Cox DR. Statistical significance tests. Br J Clin Pharmacol 1982;14:325-331.
70. Royall RM. The effect of sample size on the meaning of statistical tests. The American Statistician 1986;40:313-315.
71. Kempthorne O. Of what use are tests of sinificnace and tests of hypotheses. Communications in Statistics-Theory and Methods A5 1976;8:763-777.
72. Cox DR. Another comment on the role of statistical methods. BMJ 2001;322:231.
73. Berger VW. In defense of hypothesis testing. Br Med J 2001;Rapid response 9 September @ http://bmj.bmjjournals.com/cgi/eletters/322/7295/1184/a#16449.
74. Dracup C. Hypothesis testing-What it really is. The Psychologist 1995;8:359-362.
75. Smith SM. Clarifying the evidence. BMJ 2001;Rapid response: 28 January @ http://bmj.bmjjournals.com/cgi/eletters/322/7280/226#12264
76. Cornfield J. Sequential trials, sequential analysis and the likelihood principle. The American Statistician 2004;20:18-23.
77. Bandt CL, Boen JR. A prevalent misconception about sample size, statistical significance, and clinical importance. J Periodontol 1972;43:181-183.
78. Freeman PR. The role of p-values in analysing trial results. Stat Med 1993;12:1443-1452.
79. Pratt JW. Review of "Testing statistical hypotheses" 1959. E.L. Lehman. Journal of the American Statistical Association 1961;56:163-167.
80. Kempthorne O. Theories of inference and data analysis. In: Bancroft TA. Statistical Papers in honour of George W Snedecor. Ames, Iowa: The Iowa State Univerity Press; 1972: 167-191.
81. Perneger TV. Sifting the evidence. Likelihood ratios are alternatives to P values. Br Med J 2001;322:1184-1185.
82. Goodman SN, Royall R. Evidence and scientific research. Am J Public Health 1988;78:1568-1574.
83. Berger JO, Sellke T. Testing a point null hypothesis: The irreconcilability of P values and evidence. Journal of the American Statistical Association 1987;82:112-122.
84. Casella G, Berger RL. Reconciling Bayesian and frequentist evidence in the one-sided testing problem. Journal of the American Statistical Association 1987;82:106-111.
85. Casella G, Berger RL. Comment on Testing precise hypotheses by J.O. Berger and M. Deampdy. Statistical Science 1987;2:344-347.
86. Meehl PE. Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. Journal of Consulting & Clinical Psychology 1978;46:806-834.
87. Sellke T, Bayarri MJ, Berger JO. Calibration of p values for testing precise null hypotheses. The American Statistician 2004;55:62-71.
88. Goodman SN. A comment on replication, p-values and evidence. Stat Med 1992;11:875-879.
89. Greenwald AG, Gonzalez R, Harris RJ, Guthrie D. Effect sizes and p values: what should be reported and what should be replicated? Psychophysiology 1996;33:175-183.
90. Posavac EJ. Using p Values to Estimate the Probability of a Statistically Significant Replication. Understanding Statistics 2002;1:101-112.
91. May K. A Note on the Use of Confidence Intervals. Understanding Statistics 2003;2:133-135.
92. Lindley DV. Discussion of 'Tests of significance in theory and practice' by D.J. Johnstone. The Statistician 1986;35:502-504.

93. Braitman LE. Confidence intervals assess both clinical significance and statistical significance. Ann Intern Med 1991;114:515-517.

94. Simon R. Why confidence intervals are useful tools in clinical therapeutics. J Biopharm Stat 1993;3:243-248.

95. Thompson WD. Statistical criteria in the interpretation of epidemiologic data. Am J Public Health 1987;77:191-194.

96. Poole C. Low P-values or narrow confidence intervals: which are more durable? Epidemiology 2001;12:291-294.

97. Feinstein AR. P-values and confidence intervals: two sides of the same unsatisfactory coin. J Clin Epidemiol 1998;51:355-360.

98. Poole C. Beyond the confidence interval. Am J Public Health 1987;77:195-199.

99. Foxman B, Frerichs RR. Epidemiology of urinary tract infection: I. Diaphragm use and sexual intercourse. Am J Public Health 1985;75:1308-1313.

100. Fleiss JL. Confidence intervals vs significance tests: quantitative interpretation. Am J Public Health 1986;76:587-588.

101. Foxman B, Frerichs RR. Response from Drs Foxman and Frerichs. Am J Public Health 1986;76:587.

102. Cutler SJ, Greenhouse SW, Cornfield, J, Schneiderman, M A. The role of hypothesis testing in clinical trials. Biometrics seminar. J Chronic Dis 1966;19:857-882.

103. Borenstein M. The case for confidence intervals in controlled clinical trials. Control Clin Trials 1994;15:411-428.

104. Senn S. Discussion of the role of P-values in analysing trial results by PR Freeman. Stat Med 1993;12:1453-1457.

105. Hughes MD. Reporting Bayesian analyses of clinical trials. Stat Med 1993;12:1651-1663.

106. Efron B. Why isn't everyone a Bayesian. The American Statistician 1986;40:1-11.

107. Fleiss JL. Significance tests have a role in epidemiologic research: reactions to A. M. Walker. Am J Public Health 1986;76:559-560.