Indeed, Berkson’s paper hints at such an approach, in the section discussing the distinction between rejecting the null hypothesis $H_0$ and accepting the alternative hypothesis $H_1$. Certainly, the Bayesian approach can be used to demonstrate that conventionally statistically significant $P$-values do not necessarily correspond to strong evidence against the null hypothesis, in the sense that the when testing a normal mean a $P$-value of 0.05 implies that the posterior probability of the null is at least 0.3 in very general circumstances. Whether the more widespread use of the Bayesian approach to statistical inference would itself reduce the misinterpretation of tests of evidence against null hypotheses remains unclear: for the reasons outlined above researchers will continue to require some form of probability statement quantifying the evidence regarding particular null hypotheses of interest. The development of Bayesian methods for model choice is an area of active research interest.

Other examples used by Berkson are of situations in which confusion is caused by non-standard hypothesis tests: these would not be controversial in modern medical statistics but rather illustrate problems of interpretation familiar to anyone attempting to teach students how to fit linear terms in regression models and to check whether their models are appropriate. Such students often share Berkson’s apparent confusion over the significant departure from linearity in his Chart 1. A first null hypothesis, of no linear association between facet number and temperature, corresponds to the slope of the regression line being equal to zero. Visual inspection of Chart 1 suggests to us, as it did to Berkson, that the data provide strong evidence against this null hypothesis. A second null hypothesis states that there is no extra-linear variation. These days, we could display the data rather better by including confidence intervals around the closed circles representing the mean facet number at each temperature, hence avoiding the misleading impression gained by Berkson that ‘it appears as straight a line as one can expect to find in biological material’. We might proceed to test the second null hypothesis by including powers of temperature in our regression model, and would no doubt reach the same conclusion as Fisher, that the data provide evidence of extra-linear variation. Finally, we might test a number of further null hypotheses to check model assumptions, such as that the errors are not heteroscedastic.

Berkson concludes by calling for ‘investigation into the finding of middle $P$’s’, although he is ‘not ready to say what this should be or just what it might mean’. It is tempting to imagine a contemporary commission of inquiry whose focus is not on the meaning of middle $P$’s, since as illustrated earlier this is well understood, but on whether blame for the continuing misinterpretation of $P$-values and significance tests lies with teachers of statistics, its students and practitioners, or the counterintuitive and difficult nature of the subject matter. I think we should be told.

References
From the perspective of Berkson’s 1942 paper\(^1\) we can look both ways—backward to the heady ferment of ideas in the four decades since Pearson pushed out the statistics boat into uncharted biological waters, or forward to the decades in which statistical thinking was heavily influenced by war-time demand for industrial utility and operational effectiveness, and then to the later decades in which not so much thinking as practice was free to blossom with the speed of electronic computation.

**1901–1942: Formative years**

Berkson’s paper is a polemic that takes lively issue with the dominant or Fisherian school of statistics that Berkson saw as teaching that the concept of a null hypothesis should dominate not merely statistics but experimental science as a whole. Berkson’s thesis is that, if any sort of hypothesis should do that, it is the Neyman-Pearsonian concept of an alternative hypothesis. The novel principle that Berkson put forward in expression of his thesis is one that squeezes more meaning out of a thesis. The novel principle that Berkson put forward in expression of his thesis is one that squeezes more meaning out of a P-value than the Fisherian canon manages to do. It has the following components:

1. Treat as erroneous Fisher’s claim that the null hypothesis is never proved or established, but is possibly disproved, in the course of experimentation (words that preceded the quotation of Fisher that Berkson put at the end of his first paragraph).
2. Consider whether there is any alternative hypothesis under which whatever value of \(P\) emerges from the test of significance would be relatively frequent (i.e. neither small nor near unity).
3. If so, treat the occurrence (rather than the value) of \(P\) as evidence in favour of the alternative hypothesis.
4. If the \(P\)-value is not frequent under the null (i.e. small or, for some significance tests, close to unity) and the alternative contradicts the null, take the \(P\)-value itself as evidence in disfavour of the null hypothesis.

With its emphasis on alternative hypotheses, it is difficult to conceive of a greater challenge to the Fisherian school, but Berkson went even further with:

> How blind is the procedure of doing some test of significance, when there is no knowledge at hand as to whether it is likely to show a significant result or not...

Berkson’s paper consists of six salvos fired into the Fisherian camp—each based on a different statistical problem. Fisher responded,\(^2\) reproduced here, to only one of these salvos—which leads one to wonder why the one and why not the other five?

Fisher always had a straightforwardly simple view of significance testing, as just one element of scientific investigation. It was for him an activity with a natural and theoretically influential order to its four components: (1) Data feature of potential scientific interest \(\rightarrow\) (2) Representation of the feature as a test statistic \(S\) \(\rightarrow\) (3) Sceptical null hypothesis \(\rightarrow\) (4) \(P\)-value as the probability under the null hypothesis that \(S\) equals or exceeds its observed value. Off-stage, alternative hypotheses may play a role as inchoate concepts in the perception of the data feature, or as components of any scientific ‘afterthought’ or development.

In Section 7 of The Design of Experiments, Fisher had written:

> It is usual and convenient for experimenters to take 5 per cent as a standard level of significance, in the sense that they are prepared to ignore all results which fail to reach this standard, and, by this means, to eliminate from further discussion the greater part of the fluctuations which chance causes have introduced into their experimental results.

A later, less circumspect, expression of this (and more) is to be found in Fisher’s response\(^3\) to the psychologist WE Hick, who had asked him *inter alia* why anyone should bother about the second tail of a two-tailed test:

> ... they [Neyman & Pearson] approach the problem entirely from the wrong end, i.e. not from the point of view of a research worker, with a basis of well grounded knowledge on which a very fluctuating population of conjectures and incoherent observations is continually under examination. In these circumstances the experimenter does know what observation it is that attracts his attention. What he needs is a confident answer to the question ‘Ought I to take any notice of that?’ This question can, of course, and for refinement of thought should, be framed as ‘Is this particular hypothesis overthrown, and if so at what level of significance, by this particular body of observations?’ It can be put in this form unequivocally only because the genuine experimenter already has the answers to all the questions that the followers of Neyman and Pearson attempt, I think vainly, to answer by merely mathematical consideration.

In other words, significance testing was for Fisher a screening device that either can protect the experimenter (or observer) from following chimeras created by chance variation from an uninteresting (null) hypothesis \(H_0\) or can positively signpost directions in which scientific gold may be discovered. The test statistic horse \(S\) has to go before any alternative hypothesis cart \(H_1\)—even before thinking about the relevant null hypothesis. It is \(S\) that attracts the experimenter’s attention or that, from the Latin roots of the word significance, ‘makes the sign’. It is what engenders the concept of one or more alternatives \(H_1\) as whatever in Nature would lead to larger values of \(S\) than would be expected when calibrated by the chance hypothesis \(H_0\). If the \(P\)-value resulting from this calibration is small enough, the experimenter is encouraged to follow the signpost and develop the science that it may point to—but must also *know how to conduct an experiment which will rarely fail to give ... a statistically significant result* (quoting Section 7 of Design of Experiments again).

We are now in a position to analyse and judge the strength of the five salvos (counter-examples to the Fisherian doctrine) to which Fisher did not return fire, and to comment somewhat presumptuously on the one where he did.

**Salvo of the 100-faced die ...**

Berkson embeds a questionably normal random sample in a larger data set—by adjoining to it the outcome of a single throw of an unquestionably fair die. The test statistic at issue is the (rather degenerate) number of black faces (0 or 1) in the outcome of the throw. In terms of Neyman-Pearsonian theory, this...
test would have ‘zero power’ in the sense that skewness of the random sample would not increase the probability of rejection of the null hypothesis of normality, and the test can therefore be dismissed as useless. Berkson is implying that, by rejecting Neyman-Pearsonian concepts, Fisher would have no grounds for not using the test and would be obliged to conclude, from one black face, that the sample was skew!

Had he responded to this salvo, Fisher would probably have retaliated with the pungency he expressed in the letter to Hick already quoted:

… the practical experimenter does not often put up a damn-fool test of significance but it is a labour of many years and much art for the ‘Theory of Testing Hypotheses’ to avoid such tests.

This would not have been bombast. In the mid-1930s Fisher had already developed two ideas more fundamental than Neyman-Pearsonian power, in order to maximize the informativeness of both statistical estimation and significance testing: sufficiency and ancillarity. Applied to the die example, the random sample itself, excluding the outcome of the throw, is a sufficient statistic for the normality question, since the (conditional) distribution of that outcome (given the random sample values) does not depend at all on the shape of the distribution that generated the random sample. The sufficiency principle—Use no more than a sufficient statistic—then excludes the die from any consideration. Exclusion is also assured if we apply the ancillarity principle: the throw outcome is an ancillary statistic because its (unconditional) distribution does not depend at all on the shape of the sample distribution, and the principle requires that we should use the (conditional) distribution of the random sample (given the ancillary) as the basis of any inference (here a significance test of normality).

… of the problem of middling $P$-values for tests of Poissonianity

Most test statistics are formulated so that interest in them is aroused only when they take values that are large, so that largeness corresponds to small $P$-values. An interesting exception to this is $\chi^2$ for which $P$-values close to unity may indicate an excessive agreement between observed and expected frequencies due to fabricated data or other causes of correlated variation. The standard $\chi^2$ test for Poissonianity discussed by Berkson exemplifies this, with small and large $P$-values corresponding to the alternatives of super-Poisson and sub-Poisson distributions respectively. Is Berkson suggesting that, because we (and Fisher) do not always need explicitly small $P$-values to reject a null hypothesis, this somehow weights the case against the Fisherian doctrine? A large $P$-value for the $\chi^2$ statistic corresponds to a small one for the reciprocal statistic, which could just as well be the one that catches the experimenter’s eye (in the Fisherian version of scientific investigation). I have tried but failed to see how this example invests middling $P$-values with any novelty, since such values are simply the intersection of the non-significant values for two test statistics that would not be brought into simultaneous activity by any data set. Is this salvo founded on no more than a distinction without a difference?—a ubiquitous feature of academic discourse that may have affected the next salvo too.

… of ‘Student’ and his haemocytometer

Here Berkson attacked Fisher’s 1935 erroneous principle, as he saw it, that the null hypothesis is never proved, but is possibly disproved. Berkson claimed that statisticians with real problems do interpret a middle $P$ as positive support for the null hypothesis. He went on to applaud ‘Student’ for actually saying (in his 1907 haemocytometer paper) that:

the probabilities 0.04, 0.68, 0.25, and 0.64, though not particularly high, are not at all unlikely in four trials, supposing our theoretical law [Poissonianity] to hold, and we are not likely to be very far wrong in assuming it to do so.

In Section 15 of the 1930 edition of Statistical Methods for Research Workers, Fisher looked at the set of haemocytometer counts that gave ‘Student’ the $P$-value of 0.64 and commented that the expected frequencies … agree well with those observed. In Section 20 (The $\chi^2$ distribution) Fisher also wrote:

The term Goodness of Fit has caused some to fall into the fallacy of believing that the higher the value of $P$ the more satisfactorily is the hypothesis verified. Values over .999 have been reported which, if the hypothesis were true, would only occur once in a thousand trials…. In these cases the hypothesis considered is as definitely disproved as if $P$ had been .001.

The agree well with and verified in these quotations suggest that for $\chi^2$ tests of goodness of fit Fisher was not far from being able to agree with Berkson and ‘accept’ in some sense an unrejected null hypothesis—one that could, after all, be determined by increasing without limit the sample size (the sum of the observed frequencies). But this harmony would not have extended so easily to significance tests in general, as the following extract from a letter3 that Fisher wrote to W Edwards Deming in 1935 makes clear:

There is a good deal in the approach chosen by Neyman and Pearson that I disagree with ... It is ... a pity that these writers have introduced the concept of ‘errors of the second kind’, i.e. of accepting an hypothesis when it is false, seeing that until the true hypothesis is specified, such errors are undefined both in magnitude and in frequency. Their phraseology also encourages the very troublesome fallacy that when a deviation is not significant the hypothesis tested should be accepted as true.

… of fetal sex

Here one may wonder whether it really was Fisher in the other camp rather than some Aunt Sally of Berkson’s imagination. Berkson was surely right to point out that the strength of evidence for the hypothesis of fetal sex non-discriminability must involve consideration of sample size of the sort that might have brought harmony with Fisher, as in Salvo Four, perhaps by invoking the likelihood function. But Berkson then uses this valid point to refute those who contend that small samples can be effectively utilised in statistical investigations when the utilization is still that of interpreting a large or middling $P$-values as evidence in favour of the unrejected null hypothesis. As the Poissonianity salvo made clear, Berkson would not have placed Fisher in such company since he there criticized Fisher for declining to interpret
such $P$-values. Had Fisher commented on this part of Berkson's paper he might have claimed that the lack of any objective technique for making allowance for sample size in the interpretation of a non-small $P$-value demonstrates the futility of Berkson's 'novel principle' and at the same time lends support to his own principle that only small $P$-values have evidential value.

... of hospital mortality rates

This salvo is the one that most clearly reveals the wide gulf between the views of our two contestants. Berkson here indicates that he has no inclination to follow the order of things that Fisher propagated: Data feature → Test statistic → Null hypothesis → $P$-value. For Berkson, test statistics appear to be items in a statistician’s repertoire that can and even should be brought out prior to inspection of data and that are equally valid for testing the null hypothesis. The statistician is almost encouraged to try out a number of tests and then think about what alternatives are particularly pointed to by specified findings with different tests. It is this approach that creates the mathematical complexity involving considerations of power that Fisher so often deplored e.g. in the letter to WE Hick already quoted which began with:

I am a little sorry that you have been worrying yourself at all with that unnecessarily portentous approach to tests of significance represented by the Neyman and Pearson critical regions, etc. ....

Those are the generalities touched on in this salvo. There is, however, a particularity of the hospital mortality data—presented as an actual experience of mortalities but without reference to any publication—that deserves comment. It is revealed when we try to reproduce what Fisher would probably have done. The principal feature of the 'data' is the apparent benefit of vac-communication for all six types of operation, which gives $P = 0.016$ by the one-tailed sign test. For this feature (which implicitly points to a class of alternative hypotheses) Fisher would surely have chosen as test statistic not the one nominated for him by Berkson, but the product of the six one-tailed 'exact' $P$-values that condition on the ancillary information in the marginal frequencies and that Fisher had presented to the Royal Statistical Society 7 years previously.4 Employing the continuity correction in ref. 5 (where one proof may also require correction!) that puts only weight half on the probability of the datum actually observed, these exact $P$-values are $0.33, 0.31, 0.49, 0.36, 0.34, 0.37$. They are all well short of individual statistical significance despite the continuity correction and together deliver a quite insignificant Fisher-combination $P$-value of $0.43$—in spite of the fact that all six exact $P$-values are less than $\frac{1}{2}$. Unless the provenance of the data is questioned, this discrepancy must be the hardiwork of pure chance—without the portentous significance that Berkson attributes to it. But Fisher might well have responded to the accusation of blindness by asking Berkson to explain the $P$-value of 0.98 in the 4th test: such close-to-unity values are usually indicative of spurious provenance.

Salvo six and Return Fire: The Drosophila eye-facets

In this example, Berkson claims that Fisher was betrayed, by adherence to the unsound principle that a small $P$-value justifies rejection of the null hypothesis, into rejecting a perfectly sound null hypothesis! The $P$-value calculated by Fisher and accepted by Berkson as germane to his argument was derived on the null hypothesis that the distribution of the number of eye-facets at any recorded temperature is adequately approximated by a normal distribution whose expectation has a straight line relationship with the temperature—and that the deviations of individual observations from expectation are independently distributed with zero mean and constant variance. Berkson does not tell us whether he had seen the paper by the biologist Hersh from which Fisher in 1924 took the eye-facet data to illustrate his new test for linearity (the technique that was later transformed by Snedecor into the now familiar $F$-test). Instead, he observes that:

a small $P$ is to be expected frequently if the regression is linear and a value of the abscissal variate, in this case the temperature, is not constant but subject to fluctuation, that on inspection it appears as straight a line as one can expect to find in biological material, and that his own judgement would be, not that the regression is nonlinear, but that the temperature has varied....

So Fisher was wrong to reject the linearity hypothesis! As it happens, the geneticist in Fisher had taken an interest in the deviations from linearity that Hersh related to a question of heterozygosity, while Hersh himself had openly conceded that the recorded temperature may have been subject to error. Fortunately, we have the pleasure of reading here Fisher’s responding salvo2—an entertaining matter after 60 years but one that, at the time, Berkson is unlikely to have welcomed.

1943–2003: Sixty years of schism and realignment

Berkson’s critique took the $P$-value as an unquestioned given. He thereby did not raise a matter that had concerned Fisher for some time, that would lead inexorably to a schism in statistical science, and that is still far from resolution. It is one that bears heavily on the question of the evidential value of significance tests. Fisher did not fully explain the basis of his concern until the 1950s when he wrote in ref. 6 that:

if we possess a unique sample … on which significance tests are to be performed, there is always … a multiplicity of populations to each of which we can legitimately regard our sample as belonging.

However, in the intervals between his innovative researches as the world’s leading statistical geneticist, Fisher did not elaborate the relative simplicity of his significance test doctrine, but concentrated on justifying his pre-war ideas of inductive fiducial inference. These were based on so-called pivotal quantities such as the error of observation $e = x - \theta$ in a single observation $x$ of a parameter $\theta$. Fisher developed a subtle line of thinking that allowed an objective and scientifically established probability distribution for $e$ to be assigned to $\theta$, as its fiducial distribution, with $x$ fixed at its observed value. (This transfer has never been widely accepted. Its mathematical structure has been sympathetically explored in ref. 7.)

Fisher moved down this road into open dispute with the Neyman/Wald/Lehmann version of statistical science that was
being developed in North America and that generated a highly mathematical theory of hypothesis testing and estimation. Both sides of an increasingly bitter controversy had to find ingenious ways of dealing with different counter-examples to their internal consistencies. Both sides were unprepared for the discovery that acceptance of the Sufficiency and Ancillarity Principles implied acceptance of the Likelihood Principle: statistical evidence lies wholly in the shape of the likelihood function determined purely by the observed data. They also became increasingly vulnerable to Bayesian onslaughts from those subscribing to one or other set of axioms of rational subjective choice. By 2003, both the Fisherian and the Neyman-Pearsonian schools can be said to have failed to achieve their initiators’ expectations (and may even have reached the buffers, some would say)—whereas the Bayesian school appears by now to be confident that it will own the professional future. Readers may find refs 9 and 10 useful in coming to their own view of the current state of affairs. What would also help would be a careful analysis of the pretension of Bayesian doctrine to identify itself with the science of statistics in its widest and most useful sense i.e. a methodology that should inspire or instill in its practitioners a resistance to unethical temptations or pressures whatever the forum in which it is invoked, ensuring that those techniques comply with the unwritten code of scientific integrity.

A personal footnote

It should be clear from my comments that I do not think that Berkson came even close to winning the arguments with Fisher. Given the stature of his opponent, is that surprising? Berkson’s paper is a fine example of the art of provocation that some find refus 9 and 10 useful in coming to their own view of the current state of affairs. What would also help would be a careful analysis of the pretension of Bayesian doctrine to identify itself with the science of statistics in its widest and most useful sense i.e. a methodology that should inspire or instill in its practitioners a resistance to unethical temptations or pressures whatever the forum in which it is invoked, ensuring that those techniques comply with the unwritten code of scientific integrity.

members of my present audience will know from their own personal and professional experience that it is to the statistician that the present age turns for what is most essential in all its more important activities. They are the ‘backroom boys’ of every significant enterprise. Sadly, many ‘backrooms’ in Britain are now subject to a different sort of expansion—not of statisticians but of ‘overseers’ (an influential mix of academics manqués, accountants, auditors, ... and so on through the alphabet) whose purposes are often indifferent to the high standards of truthful inquiry that Fisher had in mind in 1952.

Since hope springs eternal, I would like to end on a happier note. My own backroom, now free of overseers, has seen the completion of some analyses of several years’ experimental data on haemopoietic stem cells from Dr Martin Rosendaal’s University College London laboratory mice. The analyses are little more than a combination of exploratory data analysis and simple significance tests based on Fisherian techniques. But together they point to a demographic explanation of the principal feature of the data, namely, an enhancement of blood-cell repopulation by stem cell grafts of a heterozygous genotype—a finding that echoes the feature of Hersh’s eye-facet data that probably caught the eye of the geneticist in Fisher.

References