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

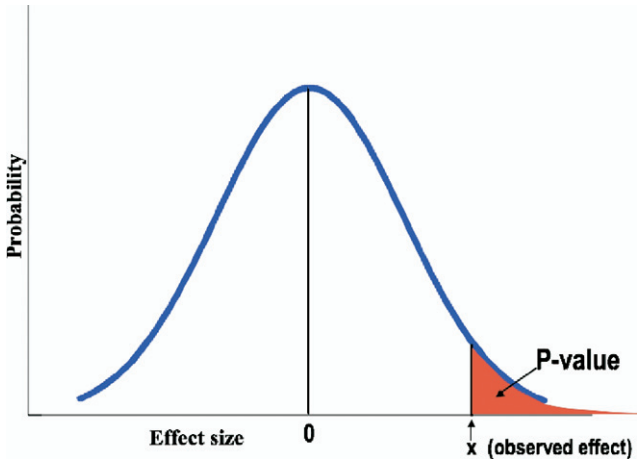

Figure 1 Graphical depiction of the definition of a (one-sided) *P* value. The curve represents the probability of every observed outcome under the null hypothesis. The *P* value is the probability of the observed outcome (x) plus all "more extreme" outcomes, represented by the shaded "tail area."

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.*

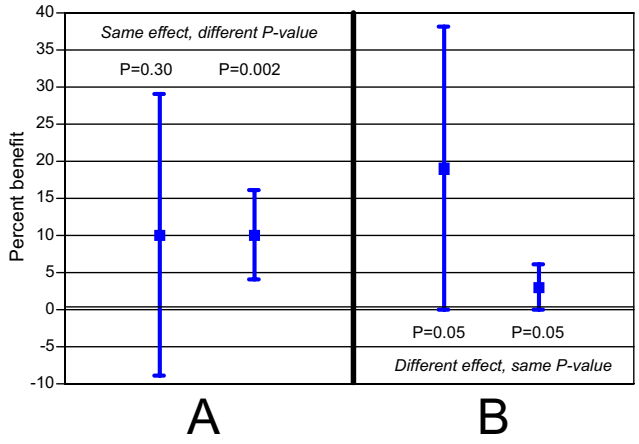

Figure 2 Figure showing how the *P* values of very different significance can arise from trials showing the identical effect with different precision (A, Misconception #4), or how same *P* value can be derived from profoundly different results (B, Misconception #5).

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

The Fisher, Neyman–Pearson Theories of Testing Hypotheses: One Theory or Two?

E. L. LEHMANN*

The Fisher and Neyman–Pearson approaches to testing statistical hypotheses are compared with respect to their attitudes to the interpretation of the outcome, to power, to conditioning, and to the use of fixed significance levels. It is argued that despite basic philosophical differences, in their main practical aspects the two theories are complementary rather than contradictory and that a unified approach is possible that combines the best features of both. As applications, the controversies about the Behrens–Fisher problem and the comparison of two binomials (2×2 tables) are considered from the present point of view.

KEY WORDS: Behrens–Fisher problem; Conditioning; Power; p-value; Significance level.

1. INTRODUCTION

The formulation and philosophy of hypothesis testing as we know it today was largely created in the period 1915–1933 by three men: R. A. Fisher (1890–1962), J. Neyman (1894–1981), and E. S. Pearson (1895–1980). Since then it has expanded into one of the most widely used quantitative methodologies, and has found its way into nearly all areas of human endeavor. It is a fairly commonly held view that the theories due to Fisher on the one hand, and to Neyman and Pearson on the other, are quite distinct. This is reflected in the fact that separate terms are often used (although somewhat inconsistently) to designate the two approaches: significance testing for Fisher and hypothesis testing for Neyman and Pearson. (Since both are concerned with the testing of hypotheses, it is convenient here to ignore this terminological distinction and to use the term "hypothesis testing" regardless of whether the testing is carried out in a Fisherian or Neyman–Pearsonian mode.)

There clearly are important differences, both in philosophy and in the treatment of specific problems. These were fiercely debated by Fisher and Neyman in a way described by Zabell (1992) as "a battle which had a largely destructive effect on the statistical profession." I believe that the ferocity of the rhetoric has created an exaggerated impression of irreconcilability. The purpose of this article is to see whether there exists a common ground that permits a resolution of some of the principal differences and a basis for rational discussion of the remaining ones.

Some of the Fisher-Neyman debate is concerned with issues studied in depth by philosophers of science (see, for example, Braithwaite 1953; Hacking 1965; Kyburg 1974; and Seidenfeld 1979). I am not a philosopher, and this article is written from a statistical, not a philosophical, point of view.

Section 2 presents some historical background for the two points of view. Section 3 discusses the basic philosophical difference between Fisher and Neyman. (Although the main substantive papers [NP 1928 and 1933a] were joint by Neyman and Pearson, their collaboration stopped soon after Neyman left Pearson's Department to set up his own program in Berkeley. After that, the debate was carried on primarily by Fisher and Neyman.) Sections 4, 5, and 6 discuss three specific issues on which the two schools differ (fixed levels versus p values, power, and conditioning). Section 7 illustrates the effect of these differences on the treatment of two statistical problems, the 2×2 table and the Behrens–Fisher problem, that have become focal points of the controversy. Finally, Section 8 suggests a unified point of view that does not resolve all questions but provides a common basis for discussing the remaining issues.

For the sake of completeness, it should be said that in addition to the Fisher and Neyman–Pearson theories there exist other philosophies of testing, of which we shall mention only two. There is Bayesian hypothesis testing, which, on the basis of stronger assumptions, permits assigning probabilities to the various hypotheses being considered. All three authors were very hostile to this formulation and were in fact motivated in their work by a desire to rid hypothesis testing of the need to assume a prior distribution over the available hypotheses.

Finally, in certain important situations tests can be obtained by an approach also due to Fisher for which he used the term *fiducial.* Most comparisons of Fisher's work on hypothesis testing with that of Neyman and Pearson (see, for example, Barnett 1982; Carlson 1976; Morrison and Henkel 1970; Spielman 1974, 1978; Steger 1971) do not include a discussion of the fiducial argument, which most statisticians have found difficult to follow. Although Fisher himself viewed fiducial considerations to be a very important part of his statistical thinking, this topic can be split off from other aspects of his work, and here I shall consider neither the fiducial nor the Bayesian approach any further.

Critical discussion of the issues considered in this article with references to the extensive literature, in a wider context and from viewpoints differing from that presented here, can be found in, for example, Oakes (1986) and Gigerenzer et al. (1989).

2. TESTING STATISTICAL HYPOTHESES

The modern theory of testing hypotheses began with Student's discovery of the t test in 1908. This was followed by

* E. L. Lehmann is Professor Emeritus, Department of Statistics, University of California, Berkeley, CA 94720. This research was supported by National Science Foundation Grant DMS-8908670. The author thanks the referees for helpful suggestions, Sandy Zabell for suggesting improvement to an early version of the article, and his wife Juliet Shaffer for discussions and critical comments at all stages.

Journal of the American Statistical Association
December 1993, Vol. 88, No. 424, Theory and Methods

Fisher with a series of papers culminating in his book *Statistical Methods for Research Workers* (1925), in which he created a new paradigm for hypothesis testing. He greatly extended the applicability of the t test (to the two-sample problem and the testing of regression coefficients) and generalized it to the testing of hypotheses in the analysis of variance. He advocated 5% as the standard level (with 1% as a more stringent alternative); through applying this new methodology to a variety of practical examples, he established it as a highly popular statistical approach for many fields of science.

A question that Fisher did not raise was the origin of his test statistics: Why these rather than some others? This is the question that Neyman and Pearson considered and which (after some preliminary work in Neyman and Pearson 1928) they later answered (Neyman and Pearson 1933a). Their solution involved not only the hypothesis but also a class of possible alternatives and the probabilities of two kinds of error: false rejection (Error I) and false acceptance (Error II). The "best" test was one that minimized P_A (Error II) subject to a bound on P_H (Error I), the latter being the significance level of the test. They completely solved this problem for the case of testing a simple (i.e., single distribution) hypothesis against a simple alternative by means of the Neyman–Pearson lemma. For more complex situations, the theory required additional concepts, and working out the details of this program was an important concern of mathematical statistics in the following decades.

The Neyman–Pearson introduction to the two kinds of error contained a brief statement that was to become the focus of much later debate. "Without hoping to know whether each separate hypothesis is true or false", the authors wrote, "we may search for rules to govern our behavior with regard to them, in following which we insure that, in the long run of experience, we shall not be too often wrong." And in this and the following paragraph they refer to a test (i.e., a rule to reject or accept the hypothesis) as "a rule of behavior".

3. INDUCTIVE INFERENCE VERSUS INDUCTIVE BEHAVIOR

Fisher considered statistics, the science of uncertain inference, able to provide a key to the long-debated problem of induction. He started one paper (Fisher 1932, p. 257) with the statement "Logicians have long distinguished two modes of human reasoning, under the respective names of deductive and inductive reasoning. . . . In inductive reasoning we attempt to argue from the particular, which is typically a body of observational material, to the general, which is typically a theory applicable to future experience." He developed his ideas in more detail in a later paper (Fisher 1935a, p. 39)

> . . . everyone who does habitually attempt the difficult task of making sense of figures is, in fact, essaying a logical process of the kind we call inductive, in that he is attempting to draw inferences from the particular to the general. Such inferences we recognize to be uncertain inferences. . . .

He continued in the next paragraph:

> Although some uncertain inferences can be rigorously expressed in terms of mathematical probability, it does not follow that mathematical probability is an adequate concept for the rigorous expression of uncertain inferences of every kind. . . . The inferences of the classical theory of probability are all deductive in character. They are statements about the behaviour of individuals, or samples, or sequences of samples, drawn from populations which are fully known. . . . More generally, however, a mathematical quantity of a different kind, which I have termed mathematical likelihood, appears to take its place [i.e., the place of probability] as a measure of rational belief when we are reasoning from the sample to the population.

Neyman did not believe in the need for a special inductive logic but felt that the usual processes of deductive thinking should suffice. More specifically, he had no use for Fisher's idea of likelihood. In his discussion of Fisher's 1935 paper (Neyman, 1935, p. 74, 75) he expressed the thought that it should be possible "to construct a theory of mathematical statistics . . . based solely upon the theory of probability," and went on to suggest that the basis for such a theory can be provided by "the conception of frequency of errors in judgment." This was the approach that he and Pearson had earlier described as "inductive behavior"; in the case of hypothesis testing, the behavior consisted of either rejecting the hypothesis or (provisionally) accepting it.

Both Neyman and Fisher considered the distinction between "inductive behavior" and "inductive inference" to lie at the center of their disagreement. In fact, in writing retrospectively about the dispute, Neyman (1961, p. 142) said that "the subject of the dispute may be symbolized by the opposing terms "inductive reasoning" and "inductive behavior." How strongly Fisher felt about this distinction is indicated by his statement in Fisher (1973, p. 7) that "there is something horrifying in the ideological movement represented by the doctrine that reasoning, properly speaking, cannot be applied to empirical data to lead to inferences valid in the real world."

4. FIXED LEVELS VERSUS p VALUES

A distinction frequently made between the approaches of Fisher and Neyman–Pearson is that in the latter the test is carried out at a fixed level, whereas the principal outcome of the former is the statement of a p value that may or may not be followed by a pronouncement concerning significance of the result.

The history of this distinction is curious. Throughout the 19th century, testing was carried out rather informally. It was roughly equivalent to calculating an (approximate) p value and rejecting the hypothesis if this value appeared to be sufficiently small. These early approximate methods required only a table of the normal distribution. With the advent of exact small-sample tests, tables of χ^2, t, F, . . . were also required. Fisher, in his 1925 book and later, greatly reduced the needed tabulations by providing tables not of the distributions themselves but of selected quantiles. (For an explanation of this very influential decision by Fisher see Kendall [1963]. On the other hand Cowles and Davis [1982] argue that conventional levels of three probable errors or two standard deviations, both roughly equivalent [in the normal case] to 5% were already in place before Fisher.) These tables allow the calculation only of ranges for the p values; however, they are exactly suited for determining the

critical values at which the statistic under consideration becomes significant at a given level. As Fisher wrote in explaining the use of his χ^2 table (1946, p. 80):

> In preparing this table we have borne in mind that in practice we do not want to know the exact value of P for any observed χ^2, but, in the first place, whether or not the observed value is open to suspicion. If P is between .1 and .9, there is certainly no reason to suspect the hypothesis tested. If it is below .02, it is strongly indicated that the hypothesis fails to account for the whole of the facts. We shall not often be astray if we draw a conventional line at .05 and consider that higher values of χ^2 indicate a real discrepancy.

Similarly, he also wrote (1935, p. 13) that "it is usual and convenient for experimenters to take 5 percent as a standard level of significance, in the sense that they are prepared to ignore all results which fail to reach this standard . . ."

Fisher's views and those of some of his contemporaries are discussed in more detail by Hall and Selinger (1986).

Neyman and Pearson followed Fisher's adoption of a fixed level. In fact, Pearson (1962, p. 395) acknowledged that they were influenced by "[Fisher's] tables of 5 and 1% significance levels which lent themselves to the idea of choice, in advance of experiment, of the risk of the 'first kind of error' which the experimenter was prepared to take." He was even more outspoken in a letter to Neyman of April 28, 1978 (unpublished; in the Neyman collection of the Bancroft Library, University of California, Berkeley): "If there had not been these % tables available when you and I started work on testing statistical hypotheses in 1926, or when you were starting to talk on confidence intervals, say in 1928, how much more difficult it would have been for us! The concept of the control of 1st kind of error would not have come so readily nor your idea of following a rule of behaviour. . . . Anyway, you and I must be grateful for those two tables in the 1925 Statistical Methods for Research Workers." (For an idea of what the Neyman–Pearson theory might have looked like had it been based on p values instead of fixed levels, see Schweder 1988.)

It is interesting to note that unlike Fisher, Neyman and Pearson (1933a, p. 296) did not recommend a standard level but suggested that "how the balance [between the two kinds of error] should be struck must be left to the investigator," and (1933b, p. 497) "we attempt to adjust the balance between the risks P_{I} and P_{II} to meet the type of problem before us."

It is thus surprising that in SMSI Fisher (1973, p. 44–45) criticized the NP use of a fixed conventional level. He objected that

> the attempts that have been made to explain the cogency of tests of significance in scientific research, by reference to supposed frequencies of possible statements, based on them, being right or wrong, thus seem to miss the essential nature of such tests. A man who 'rejects' a hypothesis provisionally, as a matter of habitual practice, when the significance is 1% or higher, will certainly be mistaken in not more than 1% of such decisions. . . . However, the calculation is absurdly academic, for in fact no scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas.

The difference between the reporting of a p value or that of a statement of acceptance or rejection of the hypothesis was linked by Fisher in Fisher (1973, pp. 79–80), to the distinction between drawing conclusions or making decisions.

> The conclusions drawn from such tests constitute the steps by which the research worker gains a better understanding of his experimental material, and of the problems which it presents. . . . More recently, indeed, a considerable body of doctrine has attempted to explain, or rather to reinterpret, these tests on the basis of quite a different model, namely as means to making decisions in an acceptance procedure.

Responding to earlier versions of these and related objections by Fisher to the Neyman–Pearson formulation, Pearson (1955, p. 206) admitted that the terms "acceptance" and "rejection" were perhaps unfortunately chosen, but of his joint work with Neyman he said that "from the start we shared Professor Fisher's view that in scientific inquiry, a statistical test is 'a means of learning' " and "I would agree that some of our wording may have been chosen inadequately, but I do not think that our position in some respects was or is so very different from that which Professor Fisher himself has now reached."

The distinctions under discussion are of course related to the argument about "inductive inference" vs. "inductive behavior," but in this debate Pearson refused to participate. He concludes his response to Fisher's 1955 attack with: "Professor Fisher's final criticism concerns the use of the term 'inductive behavior'; this is Professor Neyman's field rather than mine."

5. POWER

As was mentioned in Section 2, a central consideration of the Neyman–Pearson theory is that one must specify not only the hypothesis H but also the alternatives against which it is to be tested. In terms of the alternatives, one can then define the type II error (false acceptance) and the power of the test (the rejection probability as a function of the alternative). This idea is now fairly generally accepted for its importance in assessing the chance of detecting an effect (i.e., a departure from H) when it exists, determining the sample size required to raise this chance to an acceptable level, and providing a criterion on which to base the choice of an appropriate test.

Fisher never wavered in his strong opposition to these ideas. Following are some of his objections:

1. A type II error consists in falsely accepting H, and Fisher (1935b, p.) emphasized that there is no reason for "believing that a hypothesis has been proved to be true merely because it is not contradicted by the available facts." This is of course correct, but it does not diminish the usefulness of power calculations.

2. A second point Fisher raised is, in modern terminology, that the power cannot be calculated because it depends on the unknown alternative. For example (Fisher 1955, p. 73), he wrote:

> The frequency of the 1st class [type I error] . . . is calculable and therefore controllable simply from the specification of the null hypothesis. The frequency of the 2nd kind must depend . . . greatly on how closely they [rival hypotheses] resemble the null

> hypothesis. Such errors are therefore incalculable . . . merely from the specification of the null hypothesis, and would never have came into consideration in the theory only of tests of significance, had the logic of such tests not been confused with that of acceptance procedures. (He discussed the same point in Fisher 1947, p. 16–17.)

Fisher was of course aware of the importance of power, as is clear from the following remarks (1947, p. 24): "With respect to the refinements of technique, we have seen above that these contribute nothing to the validity of the experiment and of the test of significance by which we determine its result. They may, however, be important, and even essential, in permitting the phenomenon under test to manifest itself." The section in which this statement appears is tellingly entitled "Qualitative Methods of Increasing Sensitiveness." Fisher accepted the importance of the concept but denied the possibility of assessing it quantitatively.

Later in the same book Fisher made a very similar distinction regarding the choice of test. Under the heading "Multiplicity of Tests of the Same Hypothesis," he devoted a section (sec. 61) to this topic. Here again, without using the term, he referred to alternatives when he wrote (Fisher 1947, p. 182) that "we may now observe that the same data may contradict the hypothesis in any of a number of different ways." After illustrating how different tests would be appropriate for different alternatives, he continued (p. 185):

> The notion that different tests of significance are appropriate to test different features of the same null hypothesis presents no difficulty to workers engaged in practical experimentation but has been the occasion of much theoretical discussion among statisticians. The reason for this diversity of view-point is perhaps that the experimenter is thinking in terms of observational values, and is aware of what observational discrepancy it is which interests him, and which he thinks may be statistically significant, before he inquires what test of significance, if any, is available appropriate to his needs. He is, therefore, not usually concerned with the question: To what observational feature should a test of significance be applied?

The idea that there is no need for a theory of test choice, because an experienced experimenter knows what is the appropriate test, is expressed more strongly in a letter to W. E. Hick of October 1951 (Bennett 1990, p. 144), who, in asking about "one-tail" vs. "two-tail" in χ^2, had referred to his lack of knowledge concerning "the theory of critical regions, power, etc.":

> 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. In fact, I and my pupils throughout the world would never think of using them. If I am asked to give an explicit reason for this I should say that they 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.

6. CONDITIONAL INFERENCE

While Fisher's approach to testing included no detailed consideration of power, the Neyman–Pearson approach failed to pay attention to an important concern raised by Fisher. To discuss this issue, we must begin by considering briefly the different meanings that Fisher and Neyman attach to probability.

For Neyman, the idea of probability is fairly straightforward: It represents an idealization of long-run frequency in a long sequence of repetitions under constant conditions (see, for example, Neyman 1952, p. 27; 1957, p. 9). Later (Neyman 1977), he pointed out that by the law of large numbers, this idea permits an extension: If a sequence of independent events is observed, each with probability p of success, then the long-run success frequency will be approximately p even if the events are not identical. This property adds greatly to the appeal and applicability of a frequentist probability. In particular, it is the way in which Neyman came to interpret the value of a significance level.

On the other hand, the meaning of probability is a problem with which Fisher grappled throughout his life. Not surprisingly, his views too underwent some changes. The concept at which he eventually arrived is much broader than Neyman's: "In a statement of probability, the predicand, which may be conceived as an object, as an event, or as a proposition, is asserted to be one of a set of a number, however large, of like entities of which a known proportion, P, have some relevant characteristic, not possessed by the remainder. It is further asserted that no subset of the entire set, having a different proportion, can be recognized" (Fisher 1973, p. 113). It is this last requirement, Fisher's version of the "requirement of total evidence" (Carnap 1962, sec. 45), which is particularly important to the present discussion.

Example 1 (Cox 1958). Suppose that we are concerned with the probability $P(X \le x)$, where X is normally distributed as $N(\mu, 1)$ or $N(\mu, 4)$, depending on whether the spin of a fair coin results in heads (H) or tails (T). Here the set of cases in which the coin falls heads is a recognizable subset; therefore, Fisher would not admit the statement

$$P(X \le x) = \frac{1}{2}\Phi(x - \mu) + \frac{1}{2}\Phi\left(\frac{x - \mu}{2}\right) \qquad (1)$$

as legitimate. Instead, he would have required $P(X \le x)$ to be evaluated conditionally as

$$P(X \le x \mid \mathrm{H}) = \Phi(x - \mu) \quad \text{or}$$

$$P(X \le x \mid \mathrm{T}) = \Phi\left(\frac{x - \mu}{2}\right), \qquad (2)$$

depending on the outcome of the spin.

On the other hand, Neyman would have taken (1) to provide the natural assessment of $P(X \le x)$. Despite this preference, there is nothing in the Neyman–Pearson (frequentist) approach to prevent consideration of the conditional probabilities (2). The critical issue from a frequentist viewpoint is what to consider as the relevant replications of the experiment: a sequence of observations from the same normal

distribution or a sequence of coin tosses, each followed by an observation from the appropriate normal distribution.

Consider now the problem of testing H: $\mu = 0$ against the simple alternative $\mu = 1$ on the basis of a sample $X_1, \ldots, X_n$ from the distribution (1). The Neyman–Pearson lemma would tell us to reject H when

$$\frac{1}{2}\cdot\frac{1}{\sqrt{2\pi}}e^{-\Sigma(x_i-1)^2/2}+\frac{1}{2}\frac{1}{2\sqrt{2\pi}}e^{-\Sigma(x_i-1)^2/8} \geq K\left[\frac{1}{2}\frac{1}{\sqrt{2\pi}}e^{-\Sigma x_i^2/2}+\frac{1}{2}\frac{1}{2\sqrt{2\pi}}e^{-\Sigma x_i^2/8}\right], \quad (3)$$

where K is determined so that the probability of (3) when $\mu = 0$ is equal to the specified level α.

On the other hand, a Fisherian approach would adjust the test to whether the coin falls H or T and would use the rejection region

$$\frac{1}{\sqrt{2\pi}}e^{-\Sigma(x_i-1)^2/2} \geq K_1\frac{1}{\sqrt{2\pi}}e^{-\Sigma x_i^2/2}$$

when the coin falls H (4)

and

$$\frac{1}{2\sqrt{2\pi}}e^{-\Sigma(x_i-1)^2/8} \geq K_2\frac{1}{2\sqrt{2\pi}}e^{-\Sigma x_i^2/8}$$

when the coin falls T, (5)

where K_1 and K_2 are determined so that the null probability of both (4) and (5) is equal to α. It is easily seen that these two tests are not equivalent. Which one should we prefer?

Test (3) has the advantage of being more powerful in the sense that when the full experiment of spinning a coin and then taking n observations on X is repeated many times, and when $\mu = 1$, this test will reject the hypothesis more frequently.

The second test has the advantage that its conditional level given the outcome of the spin is α both when the outcome is H and when it is T. [The conditional level of the first test will be $<\alpha$ for one of the two outcomes and $>\alpha$ for the other.]

Which of these considerations is more important depends on the circumstances. Echoing Fisher, we might say that we prefer (1) in an acceptance sampling situation where interest focuses not on the individual cases but on the long-run frequency of errors, but that we would prefer the second test in a scientific situation where long-run considerations are irrelevant and only the circumstances at hand (i.e., H or T) matter. As Fisher put it (1973, p. 101–102), referring to a different but similar situation: "It is then obvious at the time that the judgment of significance has been decided not by the evidence of the sample, but by the throw of a coin. It is not obvious how the research worker is to be made to forget this circumstance, and it is certain that he ought not to forget it, if he is concerned to assess the weight only of objective observational facts against the hypothesis in question."

The present example is of course artificial, but the same issue arises whenever there exists an ancillary statistic (see, for example, Cox and Hinkley 1974; Lehmann 1986), and it seems to lie at the heart of the cases in which the two theories disagree on specific tests. The two most prominent of these cases are discussed in the next section.

7. TWO EXAMPLES

For many problems, a pure Fisherian or Neymann–Pearsonian approach will lead to the same test. Suppose in particular that the observations X follow a distribution from an exponential family with density

$$p_{\theta,\vartheta}(x) = C(\theta, \vartheta)e^{\theta U(x)+\Sigma_{i=1}^{k}\vartheta_i T_i(x)} \quad (6)$$

and consider testing the hypothesis

$$H: \theta = \theta_0 \quad (7)$$

against the one-sided alternatives $\theta > \theta_0$. Then Fisher would condition on $T = (T_1, \ldots, T_k)$ and would in the conditional model consider it natural to calculate the p value as the conditional probability of $U \geq u$, where u is the observed value of U. At a given level α, the result would be declared significant if $U \geq C(t)$, where $C(t)$ is determined by

$$P[U > C(t) \mid T = t] = \alpha. \quad (8)$$

A Neyman–Pearson viewpoint would lead to the same test as being uniformly most powerful among all similar tests.

But as we have seen in Example 1, the two approaches do not always lead to the same result. We next consider the two examples that have engendered the most controversy.

Example 2: The 2 × 2 table with one fixed margin. Let X, Y be two independent binomial variables with success probabilities p_1 and p_2 and corresponding to m and n trials. The problem of testing H: $p_2 = p_1$ against the alternatives $p_2 > p_1$ is of the form given by (6) and (7) with $\theta = \log[(p_2/q_2)/(p_1/q_1)]$, $T = X + Y$ and $U = Y$. Basically, there is therefore no conflict between the two approaches. However, because of the discreteness of the conditional distribution of U given t, condition (8) typically cannot be satisfied. Fisher's exact test then chooses $C(t)$ to be the largest constant for which

$$P[U > C(t) \mid T = t] \leq \alpha. \quad (9)$$

For small values of t, this may lead to conditional levels substantially less than α; for small m and n, the same may be true for the unconditional level. For this reason, Fisher's exact test has been criticized as being too conservative. Many alternatives have been proposed for which the unconditional level (which is a function of $p_1 = p_2$) is much closer to α. Upton (1982) lists 22; for other surveys, see Yates (1984) and Agresti (1992).

The issues are similar to those encountered in Example 1. If conditioning is considered appropriate (and in the present case it typically is), and if control of type I error at level α is considered essential, then the only sensible test available is of the form $U > C(t)$, where $C(t)$ is the largest value satisfying (9). If, on the other hand, only the unconditional performance is considered relevant, then we may allow the conditional level of the region $U > C(t)$ to exceed α for some values of t in such a way that the unconditional level (which is the expected value of the conditional level) gets closer to

α while remaining $\leq\alpha$ for all values of $p_2 = p_1$. This is essentially what the alternatives to Fisher's exact tests try to achieve. (The same issues arise also when analyzing 2×2 tables in which none of the margins are fixed.)

Example 3: Behrens–Fisher problem. Here we are dealing with independent samples $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ from normal distributions $N(\xi, \sigma^2)$ and $N(\eta, \tau^2)$ and we wish to test the hypothesis H: $\eta = \xi$. Against the two-sided alternatives $\eta \neq \xi$, there is general agreement that the rejection region should be of the form

$$\sqrt{\frac{|\bar{Y} - \bar{X}|}{\frac{S_X^2}{m} + \frac{S_Y^2}{n}}} \geq g\left(\frac{S_Y^2}{S_X^2}\right), \tag{10}$$

where S_Y^2 and S_X^2 are the usual estimators of σ^2 and τ^2.

Suppose that we consider it appropriate, as Fisher does, to carry out the analysis conditionally on the value of S_Y^2/S_X^2. If the conditional distribution of the left side of (10) given $S_Y^2/S_X^2 = c$ were independent of the parameters and hence known, there would be no problem. Everyone would agree to calculate g so that the conditional level is α for each c, which would then also result in an unconditional level identically equal to α. Unfortunately, the conditional distribution depends on the unknown variances. Two principal ways out of this difficulty have been proposed.

1. From a Neyman–Pearson point of view, the attempt has been to construct a function g for which the probability of (10) is $\equiv\alpha$ under H for all σ and τ (it actually depends only on the ratio $\theta = \tau^2/\sigma^2$). After much effort in this direction, it became clear that an acceptable function g satisfying this condition does not exist. But Welch and Aspin have produced tests whose level differs from α so little over the entire range of θ that, for all practical purposes, they can be viewed as solutions to the problem. (For a discussion and references see, for example, Stuart and Ord 1991, sec. 20.33.)

2. These tests are unacceptable to Fisher, however, because they admit recognizable subsets. In particular, Fisher (1956) produced an example for which the conditional level given $S_Y^2/S_X^2 = 1$ is always $>\alpha + \varepsilon$ for some positive ε. Fisher's own solution to the problem is the so-called Behrens–Fisher test, which he derived by means of a fiducial argument. Although it does not follow from this derivation, numerical evidence (Robinson 1976) strongly suggests that this test is conservative; that is, its unconditional level is always $<\alpha$. But a proof of this fact for all m, n, and θ is not yet available.

Let us call a set C in the sample space for which there exists $\varepsilon > 0$ such that

$$P_H[\text{rejecting} \mid X \in C] > \alpha + \varepsilon \quad \text{for all distributions in } H,$$

a liberally biased relevant subset. (The corresponding concept for confidence intervals is called negatively biased.) Robinson (1976) showed that no such subsets exist for the Behrens–Fisher test. (Because of this test's conservative nature, this is perhaps not too surprising.) He proposed calling a test conservative if its unconditional level is always $\leq\alpha$ and if it does not admit a liberally biased relevant subset, and expressed the hope that "perhaps the Behrens–Fisher test is optimal in some sense among the class of procedures which are conservative" (Robinson 1976, p. 970). This conjecture seems to have been disproved by Linssen (1991).

8. ONE THEORY OR TWO?

From the preceding sections it is clear that considerable differences exist between the viewpoints of Fisher and Neyman–Pearson. Are these sufficiently contradictory to preclude a unified theory that would combine the best features of both?

A first difference, discussed in Section 4, concerns the reporting of the conclusions of the analysis. Should this consist merely of a statement of significance or nonsignificance at a given level, or should a p value be reported? The original reason for fixed, standardized levels—unavailability of more detailed tables—no longer applies, and in any case reporting the p value provides more information. On the other hand, definite decisions or conclusions are often required. Additionally, in view of the enormously widespread use of testing at many different levels of sophistication, some statisticians (and journal editors) see an advantage in standardization; fortunately, this is a case where you can have your cake and eat it too. One should routinely report the p value and, where desired, combine this with a statement on significance at any stated level. (This was in fact common practice throughout the 19th Century and is the procedure frequently used by Fisher.) Two other principal differences, considered in Sections 5 and 6, are the omissions of power (by Fisher) and of conditioning (by Neyman–Pearson). It seems clear that a unified approach needs to incorporate both of these ideas.

For some problems this will cause no difficulty, because both approaches will lead to the same test, as illustrated at the beginning of Section 7. But the principles of conditioning on the one hand and of maximizing the unconditional power on the other may be in conflict, as is seen from Examples 1–3. This conflict disappears when it is realized that in such cases priority must be given to deciding on the appropriate frame of reference; that is, the real or hypothetical sequence of events that determine the meaning of any probability statement. Only after this has been settled do probabilistic concepts such as level and power acquire meaning, and it is only then that the problem of maximizing power comes into play.

This leaves the combined theory with its most difficult issue: What is the relevant frame of reference? It seems clear that even in the simplest situations (such as Ex. 1), no universal answer is possible. In any specific case, the solution will depend on contextual considerations that cannot easily be captured by a general theory.

That conflicting considerations argue for different solutions in specific cases is not an indictment of a theory, provided that the theory furnishes a basis for discussing the issues. Although Neyman and Pearson never seem to have raised the problem of just what constitutes a replication of an experiment, this question is as important for a frequentist as it is for an adherent of Fisherian probability. This was recognized, for example, by Bartlett (1984, p. 453), who

stated "I regard the 'frequence requirement of repeated sampling' as including conditional inferences." A common basis for the discussion of various conditioning concepts, such as ancillaries and relevant subsets, thus exists. The proper choice of framework is a problem needing further study.

We conclude by considering some more detailed issues and by reviewing Examples 2 and 3 from the present point of view.

1. Both Neyman–Pearson and Fisher would give at most lukewarm support to standard significance levels such as 5% or 1%. Fisher, although originally recommending the use of such levels, later strongly attacked any standard choice. Neyman–Pearson, in their original formulation of 1933, recommended a balance between the two kinds of error (i.e., between level and power). For a disucssion of how to achieve such a balance, see, for example, Sanathanan (1974). Both level and power should of course be considered conditionally whenever conditioning is deemed appropriate. Unfortunately, this is not possible at the planning stage.

2. A second point on which there appears to be no conflict between the two approaches is "truth in advertising." Even if a particular nominal level α, say 5%, is the target, when it cannot be achieved because of discreteness the test should not just be described as conservative or liberal relative to the nominal level; instead, the actual (conditional or unconditional) level should be stated. If this level is not known because it depends on unknown parameters, at least its range should be given and, if possible, also an estimated value.

3. In both the 2×2 example and the Behrens–Fisher problems, the conflict between the solutions proposed by the two schools is often discussed as that of a desire for a similar test (i.e., one for which the unconditional level is $\equiv\alpha$) versus a suitable conditional test. The issue becomes clearer if one asks for the reason that Neyman–Pearson proposed the condition of similarity. The explanation begins with the case of a simple hypothesis where these authors take it for granted that in order to maximize the power, one would want the attained level to be equal to rather than less than α. For a composite hypothesis H, they therefore stated that the level should equal α for each of the simple hypotheses making up H. The requirement for similarity thus has its origin in the desire to maximize power, the issue discussed in Section 5.

In the light of (1) and (2), a unified theory less concerned with standard nominal levels might jettison not only the demand for similarity but also that of conservatism relative to a nominal level.

When similarity cannot be achieved and conservation is not required, various compromise solutions may be available. Thus in the 2×2 case of Example 2, one could, for example, select for each t the conditional level closest to α. If this seems too permissive, then the rule could be modified by adding a cap on the conditional level beyond which one would not go. Tests with a variable conditional level that will sometimes be $<\alpha$ and sometimes $>\alpha$ have been discussed by Barnard (1989) under the name "flexible Fisher." Alternatively, one might give up on a nominal level altogether and instead for each t adjust the level to the attainable (conditional) power.

The situation is much more complicated for the Behrens–Fisher problem. On the one hand, the arguments for conditioning an S_Y^2/S_X^2 seems less compelling; on the other hand, even if this conditioning requirement is accepted, the conditional distribution depends on unknown parameters, and thus it is less clear how to control the conditional level. Robinson's formulation, mentioned in Section 7, provides an interesting possibility but requires much further investigation. But such work can be carried out from the present point of view by combining considerations of both conditioning and power.

To summarize, p values, fixed-level significance statements, conditioning, and power considerations can be combined into a unified approach. When long-term power and conditioning are in conflict, specification of the appropriate frame of reference takes priority, because it determines the meaning of the probability statements. A fundamental gap in the theory is the lack of clear principles for selecting the appropriate framework. Additional work in this area will have to come to terms with the fact that the decision in any particular situation must be based not only on abstract principles but also on contextual aspects.

[Received January 1992. Revised February 1993.]

REFERENCES

Agresti, A. (1992), "A Survey of Exact Inference for Contingency Tables" (with discussion), *Statistical Science,* 7, 131–177.

Barnard, G. A. (1989), "On Alleged Gains in Power from Lower p Values," *Statistics in Medicine,* 8, 1469–1477.

Barnett, V. (1982), *Comparative Statistical Inference* (2nd ed.), New York: John Wiley.

Bartlett, M. S. (1984), "Discussion of 'Tests of Significance for 2×2 Contingency Tables,' by F. Yates." *Journal of the Royal Statistical Society,* Ser. A, 147, 453.

Bennett, J. H. (1990), *Statistical Inference and Analysis (Selected Correspondence of R. A. Fisher),* Oxford, U.K.: Clarendon Press.

Braithwaite, R. B. (1953), *Scientific Explanation,* Cambridge, U.K.: Cambridge University Press.

Brown, L. (1967), "The Conditional Level of Student's t Test," *Annals of Mathematical Statistics,* 38, 1068–1071.

Carlson, R. (1976), "The Logic of Tests of Significance" (with discussion), *Philosophy of Science,* 43, 116–128.

Carnap, R. (1962), *Logical Foundations of Probability* (2nd ed.), Chicago: the University of Chicago Press.

Cowles, M., and Davis, C. (1982), "On the Origins of the .05 Level of Statistical Significance," *American Psychologist,* 37, 553–558.

Cox, D. R. (1958), "Some Problems Connected With Statistical Inference," *Annals of Mathematical Statistics,* 29, 357–372.

Cox, D. R., and Hinkley, D. V. (1974), *Theoretical Statistics,* London: Chapman and Hall.

Fisher, R. A. (1925) (10th ed., 1946), *Statistical Methods for Research Workers,* Edinburgh: Oliver & Boyd.

——— (1932), "Inverse Probability and the Use of Likelihood," *Proceedings of the Cambridge Philosophical Society,* 28, 257–261.

——— (1935a), "The Logic of Inductive Inference," *Journal of the Royal Statistical Society,* 98, 39–54.

——— (1935b), "Statistical Tests," *Nature,* 136, 474.

——— (1935c) (4th ed., 1947), *The Design of Experiments,* Edinburgh: Oliver & Boyd.

——— (1939), "Student," *Annals of Engenics,* 9, 1–9.

——— (1947), *The Design of Experiments* (4th ed.), New York: Hafner Press.

——— (1955), "Statistical Methods and Scientific Induction," *Journal of the Royal Statistical Society,* Ser. B, 17, 69–78.

——— (1956), "On a Test of Significance in Pearson's Biometrika Tables (No. 11), *Journal of the Royal Statistical Society,* Ser. B, 18, 56–60.

——— (1958), "The Nature of Probability," *Centennial Review,* 2, 261–274.

——— (1959), "Mathematical Probability in the Natural Sciences," *Technometrics,* 1, 21–29.
——— (1960), "Scientific Thought and the Refinement of Human Reason," *Journal of the Operations Research Society of Japan,* 3, 1–10.
——— (1973), *Statistical Methods and Scientific Inference,* (3rd ed.) London: Collins Macmillan.
Gigerenzer, G., et al. (1989), *The Empire of Chance,* New York: Cambridge University Press.
Hacking, I. (1965), *Logic of Statistical Inference,* New York: Cambridge University Press.
Hall, P., and Selinger, B. (1986), "Statistical Significance: Balancing Evidence Against Doubt," *Australian Journal of Statistics,* 28, 354–370.
Hedges, L., and Olkin, I. (1985), *Statistical Methods for Meta-Analysis,* Orlando, FL: Academic Press.
Hockberg, Y., and Tamhane, A. C. (1987), *Multiple Comparison Procedures,* New York: John Wiley.
Kendall, M. G. (1963), "Ronald Aylmer Fisher, 1890–1962," *Biometrika,* 50, 1–15.
Kyburg, H. E., Jr. (1974), *The Logical Foundations of Statistical Inference,* Boston: D. Reidel.
Linhart, H., and Zucchini, W. (1986), *Model Selection,* New York: John Wiley.
Linssen, H. N. (1991), "A Table for Solving the Behrens–Fisher Problem," *Statistics and Probability Letters,* 11, 359–363.
Miller, R. G. (1981), *Simultaneous Statistical Inference,* (2nd ed.), New York: Springer-Verlag.
Morrison, D. E., and Henkel, R. E. (1970), *The Significance Test Controversy,* Chicago: Aldine.
Neyman, J. (1935), "Discussion of Fisher (1935a)." *Journal of the Royal Statistical Society,* 98, 74–75.
——— (1938), "L'Estimation Statistique Traitée Comme un Problème Classique de Probabilité," *Actualités Scientifiques et Industrielles,* 739, 25–57.
——— (1952), *Lectures and Conferences on Mathematical Statistics and Probability* (2nd ed.), Graduate School, Washington, D.C.: U.S. Dept. of Agriculture.
——— (1955), "The Problem of Inductive Inference," *Communications in Pure and Applied Mathematics,* 8, 13–46.
——— (1956), "Note on an Article by Sir Ronald Fisher," *Journal of the Royal Statistical Society,* Ser. B, 18, 288–294.
——— (1957), " 'Inductive behavior' as a Basic Concept of Philosophy of Science," *Review of the International Statistical Institute,* 25, 7–22.
——— (1961), "Silver Jubilee of My Dispute With Fisher," *Journal of the Operations Research Society of Japan,* 3, 145–154.
——— (1966), "Behavioristic Points of View on Mathematical Statistics," in *On Political Economy and Econometrics: Essays in Honour of Oscar Lange,* Warsaw: Polish Scientific Publishers, pp. 445–462.
——— (1976), "Tests of Statistical Hypotheses and Their Use in Studies of Natural Phenomena," *Communications in Statistics, Part A—Theory and Methods,* 5, 737–751.
——— (1977), "Frequentist Probability and Frequentist Statistics," *Synthèse,* 36, 97–131.
Neyman, J., and Pearson, E. S. (1928), "On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference," *Biometrika,* 20A, 175–240, 263–294.
——— (1933a), "On the Problem of the Most Efficient Tests of Statistical Hypotheses," *Philosophical Transactions of the Royal Society of London,* Ser. A, 231, 289–337.
——— (1933b), "The Testing of Statistical Hypotheses in Relation to Probabilities A Priori," *Proceedings of the Cambridge Philosophical Society,* 29, 492–510.
Oakes, M. (1986), *Statistical Inference: A Comment for the Social and Behavioral Sciences,* New York: John Wiley.
Pearson, E. S. (1955), "Statistical Concepts in Their Relation to Reality," *Journal of the Royal Statistical Society,* Ser. B, 17, 204–207.
——— (1962), "Some Thoughts on Statistical Inference," *Annals of Mathematical Statistics,* 33, 394–403.
——— (1974), "Memories of the Impact of Fisher's Work in the 1920's," *International Statistical Review,* 42, 5–8.
Pearson, E. S., and Hartley, H. O. (1954), *Biometrika Tables for Statisticians (Table No. 11),* New York: Cambridge University Press.
Pedersen, J. G. (1978), "Fiducial Inference," *International Statistical Review,* 46, 147–170.
Robinson, G. K. (1976), "Properties of Student's t and of the Behrens–Fisher Solution to the Two-Means Problem," *The Annals of Statistics,* 4, 963–971.
——— (1982), "Behrens–Fisher Problem," in Encyclopedia of Statistical Sciences (Vol. 1, eds. S. Kotz and N. L. Johnson), New York: John Wiley, pp. 205–209.
Savage, L. J. (1976), "On Rereading R. A. Fisher" (with discussion), *Annals of Statistics,* 4, 441–500.
Schweder, T. (1988), "A Significance Version of the Basic Neyman–Pearson Theory for Scientific Hypothesis Testing," *Scandanavian Journal of Statistics,* 15, 225–242.
Seidenfeld, T. (1979), *Philosophical Problems of Statistical Inference,* Boston: D. Reidel.
Spielman, S. (1974), "The Logic of Tests of Significance," *Philosophy of Science,* 41, 211–226.
——— (1978), "Statistical Dogma and the Logic of Significance Testing," *Philosophy of Science,* 45, 120–135.
Steger, J. A. (ed.) (1971), *Readings in Statistics for the Behavioral Scientist,* New York: Holt, Rinehart and Winston.
Stuart, A., and Ord, J. K. (1991), *Kendall's Advanced Theory of Statistics, Vol. II* (5th ed.), New York: Oxford University Press.
Tukey, J. W. (1960), "Conclusions vs. Decisions," *Technometrics,* 2, 424–432.
Upton, G. J. G. (1982), "A Comparison of Alternative Tests for the 2×2 Comparative Trial," *Journal of the Royal Statistical Society,* Ser. A, 145, 86–105.
Wallace, D. L. (1980), "The Behrens–Fisher and Fieller–Creasy Problems," in *R. A. Fisher: An Application,* eds. S. E. Fienberg and D. V. Hinkley, New York: Springer-Verlag, pp. 119–147.
Yates, F. (1984), "Tests of Significance for 2×2 Contingency Tables" (with discussion), *Journal of the Royal Statistical Society,* Ser. A, 147, 426–463.
Zabell, S. L. (1992), "R. A. Fisher and the Fiducial Argument," *Statistical Science,* 7, 369–387.

Statistical Concepts in the Relation to Reality

E. S. Pearson

Journal of the Royal Statistical Society. Series B (Methodological), Vol. 17, No. 2 (1955), 204-207.

Stable URL:
http://links.jstor.org/sici?sici=0035-9246%281955%2917%3A2%3C204%3ASCITRT%3E2.0.CO%3B2-M

Journal of the Royal Statistical Society. Series B (Methodological) is currently published by Royal Statistical Society.

Statistical Concepts in their Relation to Reality

By E. S. Pearson

University College, London

SUMMARY

This paper contains a reply to some criticisms made by Sir Ronald Fisher in his recent article on "Statistical Methods and Scientific Induction".

Controversies in the field of mathematical statistics seem largely to have arisen because statisticians have been unable to agree on how theory is to provide, in terms of probability statements, the numerical measures most helpful to those who have to draw conclusions from observational data. We are concerned here with the ways in which mathematical theory may be put, as it were, into gear with the common processes of rational thought, and there seems no reason to suppose that there is one best way in which this can be done. If, therefore, Sir Ronald Fisher recapitulates and enlarges on his views upon statistical methods and scientific induction we can all only be grateful, but when he takes this opportunity to criticize the work of others through misapprehension of their views as he has done in his recent contribution to this *Journal* (Fisher 1955), it is impossible to leave him altogether unanswered.

In the first place it seems unfortunate that much of Fisher's criticism of Neyman and Pearson's approach to the testing of statistical hypotheses should be built upon a "penetrating observation" ascribed to Professor G. A. Barnard, the assumption involved in which happens to be historically incorrect. There was no question of a difference in point of view having "originated" when Neyman "re-interpreted" Fisher's early work on tests of significance "in terms of that technological and commercial apparatus which is known as an acceptance procedure". There was no sudden descent upon British soil of Russian ideas regarding the function of science in relation to technology and to five-year plans. It was really much simpler—or worse. The original heresy, as we shall see, was a Pearson one!

As has often been pointed out, the break with the traditional approach to the handling of tests for the significance of differences came with Student's paper of 1908, although the implications of the step which he had taken were not realized for some time. In puzzling over the relation to this step of Fisher's early theoretical papers and the first edition of his *Statistical Methods for Research Workers*, during the years 1925–27, I could not satisfy myself that the reasons which had been given for the choice of a particular test function in terms of the theory of estimation were altogether adequate. It was a question which I discussed from time to time with Student, and I have already quoted (Pearson, 1938, p. 243) a letter of his written in 1926 which contained the germ of that fruitful idea about the hypothesis tested and its alternatives. Apart from Student, I had no contact with industry at that time and it was some years before the publications of W. A. Shewhart appeared, showing the scope for statistical method in problems of acceptance sampling. Indeed, to dispel the picture of the Russian technological bogey, I might recall how certain early ideas came into my head as I sat on a gate overlooking an experimental blackcurrant plot at the East Malling Research Station!

To the best of my ability I was searching for a way of expressing in mathematical terms what appeared to me to be the requirements of the scientist in applying statistical tests to his data. After contact was made with Neyman in 1926, the development of a joint mathematical theory proceeded much more surely; it was not till after the main lines of this theory had taken shape with its necessary formalization* in terms of critical regions, the class of admissible hypotheses, the two sources of error, the power function, etc., that the fact that there was a remarkable parallelism of ideas in the field of acceptance sampling became apparent. Abraham Wald's contri-

* Necessary just as was the introduction of such terms as "sufficiency" and "amount of information" in the formal development of Fisher's theory.

butions to decision theory of ten to fifteen years later were perhaps strongly influenced by acceptance sampling problems, but that is another story.

So much for historical clarification. Turn now to some of Professor Fisher's strictures. It seems to me that he is often tilting at views which those whom he attacks have never held. Where, for example, do we really stand in regard to the phrase "repeated sampling from the same population"? As Fisher points out (first paragraph of his p. 71), if we have before us a single sample of observations resulting perhaps from some experimental procedure, the population of possible samples, which may be termed the reference set, will often have no objective reality, being only a product of the statistician's imagination. Further, he remarks:

> "In respect of tests of significance, therefore, there is need for further guidance as to how this imagination is to be exercised. In fact a careful choice has to be made, based on the understanding of the question or questions to be answered".

I agree entirely with this need for careful choice, but this was just what Neyman and I were pointing out long ago. The difference often appears to lie in the particular population of samples considered most appropriate. The two examples which Fisher first discusses, those of linear regression and the 2×2 table, do in fact as a pair throw much light on the question of what is involved in this exercise of the imagination.

Professor Fisher's choice of reference set is based, I think, on his theory of information. Thus in writing of the 2×2 table he speaks (p. 73) of "the reasonable principle that in testing the significance with a unique sample, we should compare it only with other possibilities in all relevant respects like that observed", and again, in the regression problem, he refers to "a population of samples in all relevant respects like that observed". The meaning of terms such as "relevant' is not of course self-evident without definition, but such phrases form part of Fisher's general approach to estimation theory and the reference sets adopted in these two examples are made perfectly clear.

We may, however, ask whether there are not other "reasonable principles" which might be used to guide the statistician's imagination. Here, for example, is one. If probability is to be justly applied to the analysis of data, it follows that a random process must have been introduced or been naturally present at some stage in the collection of these data. Is there not then an appeal to the imagination in taking as the hypothetical population of samples that which would have been generated by repetition of this random process?

If we follow this principle in the regression problem, we see that the reference set will depend on the character of the experiment or investigation. If the values of x were chosen in advance, then the population of samples consists of those having these fixed x's, but with varying y values. If the data were obtained by sampling N pairs of observations (x, y) freely from a bivariate population,* the population of samples may be imagined as enlarged accordingly. In both cases, however, $t = (b - \beta)\sqrt{A}/s$ will follow Student's distribution, although in the second case $A = S(x - \bar{x})^2$ will vary from sample to sample, as do b and s. From the Neyman and Pearson point of view, t would be regarded in both instances as the appropriate function of the sample to use in testing the hypothesis that $b = \beta_0$ and the awkwardness of the distribution of b itself in the second situation would be irrelevant. The population of samples having A fixed, which is the reference set of Professor Fisher's approach, can clearly be imagined but does not seem to have any experimental counterpart which, of course, from his point of view it need not.

The case of the 2×2 table provides an interesting companion example. Here, as Barnard first pointed out, data presented in the same form of table may have been obtained from a sampling or an experiment conducted in several different ways, For example, they may arise: (i) after the random partition of a number, N, of individuals into two groups which receive different "treatments"; (ii) by drawing randomly and independently a sample from each of two populations; (iii) by drawing a single random sample from a population of individuals possessing two qualitative characters. Following Fisher and Yates, the statistician should in each case relate his test of significance to the same reference set, that of the population of samples giving to the table the same marginal totals as those observed. The other principle to which I have referred would define three different reference sets, of which only that for case (i) corresponds with the Fisher and Yates set.

* In which the array distributions of y for fixed x are, of course, normal and homoscedastic.

Because we are dealing with discontinuous hypergeometric distributions and not with the normal curve, we do not obtain from this second approach, as in the case of linear regression, a test function whose distribution is the same for all three reference sets. All that we can say* is that if the table is denoted by

a	c	m
b	d	n
r	s	N

then under the null hypothesis

$$u = \frac{a - mr/N}{\left\{\frac{mnrs}{N^2(N-1)}\right\}^{\frac{1}{2}}}$$

will for all reference sets have zero expectation and unit variance.

But does the existence of this limitation establish that one principle is right, the other wrong? I think not, because there is still a further matter to be considered which is often overlooked. Having decided on the reference set that he regards as appropriate, Professor Fisher has still to set out the logical justification for measuring the level of significance in terms of the integral or the sum of the separate probabilities in the tail of the relevant probability distribution. This is a matter which has been raised by Harold Jeffreys (1948, p. 357) and again by G. A. Barnard (1949, p. 137). Starting from the reference set which they considered appropriate, Neyman and I arrived at the critical or rejection region for the sample point through a formulation of the alternatives to the null hypothesis, and as soon as these are considered in the problem of the 2×2 table it appears necessary to differentiate the cases (i), (ii) and (iii). Given the critical region, there is clearly more than one numerical measure which could be associated with it. We deliberately chose the integral or sum of the probabilities (under the null hypothesis) of the sample point falling within the region rather than, say, the value of the ratio of likelihoods on its boundary because it seemed to us the more relevant and meaningful measure to use.

It seems to me that there is still a good deal here that is worth thinking over and that we shall get no nearer to a solution of the logical problems involved by throwing up the question "repeated samples from the same population?" and answering, in effect, "what nonsense!" We have only to turn to D. V. Lindley's recent paper (1953) and the discussion which followed to realize the continued value of an unrestricted play of thought round these problems.

Professor Fisher's next objection is to the use of such terms as the "acceptance" or "rejection" of a statistical hypothesis, and "errors of the first and second kinds". It may be readily agreed that in the first Neyman and Pearson paper of 1928, more space might have been given to discussing how the scientific worker's attitude of mind could be related to the formal structure of the mathematical probability theory that was introduced. Nevertheless it should be clear from the first paragraph of this paper that we were not speaking of the *final* acceptance or rejection of a *scientific* hypothesis on the basis of statistical analysis. We speak of accepting or rejecting a hypothesis with a "greater or less degree of confidence". Further, we were very far from suggesting that statistical methods should force an irreversible acceptance procedure upon the experimenter. Indeed, from the start we shared Professor Fisher's view that in scientific enquiry, a statistical test is "a means of learning", for we remark: "the tests themselves give no final verdict, but as tools help the worker who is using them to form his final decision". No doubt we could more aptly have said "his final or provisional decision"; even scientists, if they are employed in research departments by industry or government, may sometimes have to give a final decision.

As already mentioned, a certain simplification of real situations and a formalization in the verbal expression of ideas seems unavoidable when one attempts to put mathematical theory into gear with the way the mind works. I would agree that some of our wording may have been chosen inadequately, but I do not think that our position in some respects was or is so very

* Unless we follow K. D. Tocher's (1950) suggestion of adding to a a random variable uniformly distributed in the interval (0, 1). Then, we can use a test function whose probability distribution is the same for all the three reference sets of cases (i), (ii) and (iii).

different from that which Professor Fisher himself has now reached. On p. 73 of his last (1955) paper, he sets out as alternatives (*a*) and (*b*) what he thinks may be, according to the circumstances, the worker's attitude in a case where the test of significance applied gives no strong reason for rejecting the null hypothesis. The phrases used, cautious though they are, are yet so relevant to the understanding of the Neyman and Pearson approach, that I shall quote them here. The worker is stated to express himself as follows:

> (*a*) "The possible deviation from truth of my working hypothesis, to examine which the test is appropriate, seems not to be of sufficient magnitude to warrant any immediate modification".

or

> (*b*) "The deviation is in the direction expected for certain influences which seemed to me not improbable, and to this extent my suspicion has been confirmed; but the body of data available so far is not by itself sufficient to demonstrate their reality".

What ideas seem to underlie these statements? There is a recognition that the probable deviations from the working hypothesis lie in a particular direction. This seems to imply that the appropriate test is one which should be sensitive, indeed as sensitive as possible, to deviations in that direction. If he is going to have to discard his working hypothesis, the scientist would presumably like to be able to reach the conclusion that this is necessary with the greatest economy of effort in experimentation. Under these conditions, as part of the mathematical structure which would help to determine the appropriate test (or to compare alternative tests), Neyman and I introduced the notions of the class of admissible hypotheses and the power function of a test. The class of admissible alternatives is formally related to the direction of probable deviations —changes in mean, changes in variability, departure from linear regression, existence of interactions, or what you will. The power function will help to indicate what amount of data may be required to demonstrate the reality of specific departures from the working hypothesis.

It seems to me that continuing on the lines of statements (*a*) and (*b*), we may imagine our worker to go further and to enlarge on the term "appropriate" as follows:

> (*c*) "The appropriate test is one which, while involving (through the choice of its significance level) only a very small risk of discarding my working hypothesis prematurely will enable me to to demonstrate with assurance (but without an unnecessary amount of experimentation) the reality of the influences which I suspect may be present".

If we accept (*c*) as a reasonable expression of attitude, it seems to follow that our worker has among other things two balancing considerations in his mind; he wants to avoid:

> (1) discarding his working hypothesis prematurely,
>
> (2) waiting an unnecessarily long time before reaching the conclusion that suspected factors are influencing the situation.

The formal description of this situation as involving the Scylla and Charybdis of two possible "sources of error", may be abhorrent to him. But perhaps, cautious as this ideal scientist is, he would admit to a desire to avoid being wrong in a tentative opinion expressed, let us say, in an informal discussion following another scientific colleague's paper read before a learned society!

Professor Fisher's final criticism concerns the use of the term "inductive behaviour"; this is Professor Neyman's field rather than mine.

References

BARNARD, G. A. (1949), *J. R. Statist. Soc.* B, **11**, 115–139.
FISHER, R. A. (1955), *J. R. Statist. Soc.* B, **17**, 69–78.
JEFFREYS, H. (1948), *Theory of Probability*. Oxford University Press.
LINDLEY, D. V. (1953), *J. R. Statist. Soc.* B, **15**, 30–76.
NEYMAN, J., & PEARSON, E. S. (1928), *Biometrika*, **20A**, 175–240.
PEARSON, E. S. (1938), *Biometrika*, **30**, 210–50.
TOCHER, K. D. (1950), *Biometrika*, **37**, 130–44.

Statistical Methods and Scientific Induction

Ronald Fisher

Journal of the Royal Statistical Society. Series B (Methodological), Vol. 17, No. 1 (1955), 69-78.

Stable URL:
http://links.jstor.org/sici?sici=0035-9246%281955%2917%3A1%3C69%3ASMASI%3E2.0.CO%3B2-M

Journal of the Royal Statistical Society. Series B (Methodological) is currently published by Royal Statistical Society.

Statistical Methods and Scientific Induction

By Sir Ronald Fisher

Department of Genetics, University of Cambridge

SUMMARY

The attempt to reinterpret the common tests of significance used in scientific research as though they constituted some kind of acceptance procedure and led to "decisions" in Wald's sense, originated in several misapprehensions and has led, apparently, to several more.

The three phrases examined here, with a view to elucidating the fallacies they embody, are:

(i) "Repeated sampling from the same population",
(ii) Errors of the "second kind",
(iii) "Inductive behaviour".

Mathematicians without personal contact with the Natural Sciences have often been misled by such phrases. The errors to which they lead are not always only numerical.

1. *Introduction*

During the present century a good deal of progress seems to have been made in the business of interpreting observational data, so as to obtain a better understanding of the real world. The three aspects of principle importance for this progress have been, first, the use of better mathematics and more comprehensive ideas in mathematical statistics; leading to more correct or exact methods of calculation, applied to the given body of data (a unique sample in the language of W. S. Gosset, writing under the name of "Student") which comprehends all the numerical information available on the topic under discussion. Secondly, as methods of summarizing and drawing correct conclusions approached adequacy, the wide subject of experimental design was opened up, aimed at obtaining data more complete and precise, and at avoiding waste of effort in the accumulation of ill-planned, indecisive, or irrelevent observations. Thirdly, as a natural or even inevitable concomitant of the first two, a more complete understanding has been reached of the structure and peculiarities of inductive logic—that is of reasoning from the sample to the population from which the sample was drawn, from consequences to causes, or in more logical terms, from the particular to the general.

Much that I have to say will not command universal assent. I know this for it is just because I find myself in disagreement with some of the modes of exposition of this new subject which have from time to time been adopted, that I have taken this opportunity of expressing a different point of view; different in particular from that expressed in numerous papers by Neyman, Pearson Wald and Bartlett. There is no difference to matter in the field of mathematical analysis, though different numerical results are arrived at, but there is a clear difference in logical point of view, and I owe to Professor Barnard of The Imperial College the penetrating observation that this difference in point of view originated when Neyman, thinking that he was correcting and improving my own early work on tests of significance, as a means to the "improvement of natural knowledge", in fact reinterpreted them in terms of that technological and commercial apparatus which is known as an acceptance procedure.

Now, acceptance procedures are of great importance in the modern world. When a large concern like the Royal Navy receives material from an engineering firm it is, I suppose, subjected to sufficiently careful inspection and testing to reduce the frequency of the acceptance of faulty or defective consignments. The instructions to the Officers carrying out the tests must also, I conceive, be intended to keep low both the cost of testing and the frequency of the rejection of satisfactory lots. Much ingenuity and skill must be exercised in making the acceptance procedure a really effectual and economical one. I am casting no contempt on acceptance procedures, and

I am thankful, whenever I travel by air, that the high level of precision and reliability required can really be achieved by such means. But the logical differences between such an operation and the work of scientific discovery by physical or biological experimentation seem to me so wide that the analogy between them is not helpful, and the identification of the two sorts of operation is decidedly misleading.

I shall hope to bring out some of the logical differences more distinctly, but there is also, I fancy, in the background an ideological difference. Russians are made familiar with the ideal that research in pure science can and should be geared to technological performance, in the comprehensive organized effort of a five-year plan for the nation. How far, within such a system, personal and individual inferences from observed facts are permissible we do not know, but it may be safer, and even, in such a political atmosphere, more agreeable, to regard one's scientific work simply as a contributary element in a great machine, and to conceal rather than to advertise the selfish and perhaps heretical aim of understanding for oneself the scientific situation. In the U.S. also the great importance of organized technology has I think made it easy to confuse the process appropriate for drawing correct conclusions, with those aimed rather at, let us say, speeding production, or saving money. There is therefore something to be gained by at least being able to think of our scientific problems in a language distinct from that of technological efficiency.

I believe I can best illustrate the contrast I want to make clear by taking a few current phrases which are foreign to my own point of view, and after examining these, by setting out in a more constructive spirit, some of the special characteristics of inductive reasoning. The phrases I should choose for the fallacies they embody are:

(i) Repeated sampling from the same population.
(ii) Errors of the "second kind".
(iii) "Inductive behaviour".

But first I must exemplify the extent to which divergence in language has been carried by quoting some rather simple phrases from Wald's book on Decision Functions.

On the outside of the cover we read, "Particularly noteworthy is the treatment of experiment design as a part of the general decision problem".

On the inside, "The design of experimentation is made a part of the general decision problems—a major advance beyond previous results", and in the first paragraph of the author's preface "A major advance beyond previous results is the treatment of the design of experimentation as a part of the general decision problem".

These claims seem very much like an afterthought, of a kind which is sometimes suggested by a publisher; for, apart from these three quotations, the design of experiments is scarcely mentioned in the rest of the book. For example, the index does not contain the word "replication", or "control", or "randomization"; there is no discussion of the functions and purposes of these three elements of design. Of authorities, the bibliography does not contain the names of Yates, of Finney, or of Davies; or, on the other side of the Atlantic, of Goulden, who was the first of transatlantic writers on the design of experiments, or of Cochran and Cox. My own book is indeed mentioned, but no use seems to have been made of it. The obvious inference is that Wald was quite unaware of the nature and scope of the subject of experimental design, but had simply assumed that it *must* be included in that of acceptance procedures, to which his book is devoted. Rather similar, equally innocent and unfounded presumptions, have been not uncommon in the last twenty years. They would scarcely have been possible without that insulation from all living contact with the natural sciences, which is a disconcerting feature of many mathematical departments.

The first questionable phrase and the one responsible for the greatest amount of *numerical* error is:

2. "*Repeated Sampling from the Same Population*"

The operative properties of an acceptance procedure, single or sequential, are ascertained practically or conceptually by applying it to a series of successive similar samples from the same source of supply, and determining the frequencies of the various possible results. It is doubtless in consequence of this that it has been thought, and frequently asserted, that the validity of a

test of significance is to be judged in the same way. However, a rather large number of examples are now known in which this rule is seen to be misleading. The root of the difficulty of carrying over the idea from the field of acceptance procedures to that of tests of significance is that, where acceptance procedures are appropriate, the source of supply has an objective reality, and the population of lots, or one or more, which could be successively chosen for examination is uniquely defined; whereas if we possess a unique sample in student's sense on which significance tests are to be performed, there is always, as Venn (1876) in particular has shown, a multiplicity of populations to each of which we can legitimately regard our sample as belonging; so that the phrase "repeated sampling from the same population" does not enable us to determine which population is to be used to define the probability level, for no one of them has objective reality, all being products of the statistician's imagination. In respect of tests of significance, therefore, there is need for further guidance as to how this imagination is to be exercised. In fact a careful choice has to be made, based on an understanding of the question or questions to be answered. By ignoring this necessity a "theory of testing hypotheses" has been produced in which a primary requirement of any competent test has been overlooked.

Consider the case of simple linear regression. Let us suppose that the numerical data consist of N pairs of values (x, y), while the qualitative data tell us that for each value of x the variate y is normally distributed with variance σ about a mean given by

$$Y \equiv E_x(y) = \alpha + x\beta, \qquad E_x(y - Y)^2 = \sigma^2,$$

being a linear function of the variate x. The qualitative data may also tell us how x is distributed, with or without specific parameters; this information is irrelevant.

In such cases the unknown parameter, β, may be estimated and the precision of estimation determined by a standard and well known procedure; let

$$A = S(x - \bar{x})^2, \quad B = S(x - \bar{x})(y - \bar{y}), \quad C = S(y - \bar{y})^2.$$

Then we may take as our estimate of β the statistic

$$b = B/A,$$

and of σ^2 the statistic

$$s^2 = (C - B^2/A) \div (N - 2).$$

For samples having the same value A it is easy to show that the estimate b is normally distributed about β with variance σ^2/A, so that we have a typical analysis of variance:

d.f.	*Sum of Squares*	*Mean Square*
1	$A(b - \beta)^2$	$A(b - \beta)^2$
$N - 2$	$C - B^2/A$	s^2

and the significance of the deviation of b from zero, or any other proposed value of β, is a simple t-test with $N - 2$ degrees of freedom, with

$$t = (b - \beta_0)\frac{\sqrt{A}}{s},$$

where β_0 is the theoretical value proposed for comparison.

I do not believe that anyone doubts the validity of this simple test. It does, however, violate the rule of determining levels of significance by frequencies of occurrence of the proposed events in repeated samples from the same population. For if a succession of sets of N pairs of observations (x, y) were taken from the same population, the value of A would not be the same for each set. Consequently, the frequency distribution of $b - \beta$ in the aggregate of all such sets would not be the same as that which I have calculated taking A constant, and would indeed be unknown until the sampling variation of A were investigated. In reality, therefore, no one uses the rule of determining the level of significance by successive sampling from the population of *all* random samples of N pairs of values, but, ever since the right approach was indicated (Fisher 1922), the *selection* of all random samples having a constant value A, equal to that actually observed in the

sample under test, is what has in fact been used. The normal distribution of b about β with variance σ^2/A does not correspond with any realistic process of sampling for acceptance, but to a population of samples in all relevant respects like that observed, neither more precise nor less precise, and which therefore we think it appropriate to select in specifying the precision of the estimate, b. In relation to the estimation of β the value A is known as an *ancillary statistic*. Had it been necessary we should not have hesitated to specify all values of x (x_1, . . . , x_N) individually, but this would have made no difference once the comprehensive value A had been specified.

The confusion introduced, even in the case of the most fundamental and logically simple of tests of significance, by the introduction of the notion of basing the test on repeated sampling from the same population, is well illustrated by some episodes, which ought not to be forgotten, in the curious history of testing proportionality in a two-by-two table.

In the solution of the problem of the 2 × 2 table, put forward concurrently by Dr. F. Yates and myself in 1934, the essential point was the recognition that the probabilities of occurrence of different possible tables, *having the same marginal totals*

a	b	$a+b$
c	d	$c+d$
$a+c$	$b+d$	n

were proportional simply to

$$1/a\,!\,b\,!\,c\,!\,d\,!$$

where a, b, c and d are the four frequencies observed in the double dichotomy, whatever might be the probabilities governing the marginal distributions. Within sets of tables having the same margins, therefore, each may be assigned an absolute probability:

$$\frac{(a+b)\,!\,(a+c)\,!\,(b+d)\,!\,(c+d)\,!}{n\,!} \cdot \frac{1}{a\,!\,b\,!\,c\,!\,d\,!}$$

where the new factor depends only on the margins and not on the contents.

In this case the margins of the table, which by themselves supply no information as to the proportionality of the contents, do, like the value A in the regression example, determine how much information the contents will contain. The reasonable principle that in testing the significance with a unique sample, we should compare it only with other possibilities in all relevant respects like that observed, will lead us to set aside the various possible tables having different margins, the relative frequencies of which must depend on unknown factors of the population sampled.

On two occasions in the intervening twenty years distinguished statisticians have attempted to bring into the account populations of fourfold tables not having fixed margins. In both cases, such is the reasonableness of human nature in favourable cases, the authors of these innovations withdrew them after some discussion, and expressed themselves as completely satisfied that the apparent advance they had made was illusory. The first was Professor E. B. Wilson of the Harvard School of Public Health, writing in *Science* in 1941, and later taking occasion to expound the method of Fisher and Yates in two papers in the *Proceedings of the National Academy of Sciences* in the following year. The second case was that of Professor Barnard, who started on the assumption that the method expounded by Neyman and Pearson could be relied on, and in the first flush of success reported a test using the language of that theory "much more powerful than Fisher's", but who also, after some discussion, had the generosity to go out of his way to explain that further meditation had led him to the conclusion that Fisher was right after all.

Professor Barnard has a keen and highly trained mathematical mind, and the fact that he was misled into much wasted effort and disappointment should be a warning that the theory of testing hypotheses set out by Neyman and Pearson has missed at least some of the essentials of the problem, and will mislead others who accept it uncritically. Indeed, in the matter of Behren's test for the significance of the difference between the means of two small samples, objection was taken on exactly the ground that the significance level is not the same as the frequency found on repeated sampling.

The examples I have given from simpler problems show clearly that it should never have been put forward in the field of significance tests, though perhaps perfectly appropriate to acceptance sampling.

3. *Errors of the "Second Kind"*

The phrase "Errors of the second kind", although apparently only a harmless piece of technical jargon, is useful as indicating the type of mental confusion in which it was coined.

In an acceptance procedure lots will sometimes be accepted which would have been rejected had they been examined fully, and other lots will have been rejected when, in this sense, they ought to have been accepted. A well-designed acceptance procedure is one which attempts to minimize the losses entailed by such events. To do this one must take account of the costliness of each type of error, if errors they should be called, and in similar terms of the costliness of the testing process; it must take account also of the frequencies of each type of event. For this reason probability *a priori*, or rather knowledge based on past experience, of the frequencies with which lots of different quality are offered, is of great importance; whereas, in scientific research, or in the process of "learning by experience", such knowledge *a priori* is almost always absent or negligible.

Simply from the point of view of an acceptance procedure, though we may by analogy think of these two kinds of events as "errors" and recognize that they are errors in opposite directions, I doubt if anyone would have thought of distinguishing them as of two kinds, for in this *milieu* they are essentially of one kind only and of equal theoretical importance. It was only when the relation between a test of significance and its corresponding null hypothesis was confused with an acceptance procedure that it seemed suitable to distinguish errors in which the hypothesis is rejected wrongly, from errors in which it is "accepted wrongly" as the phrase does. The frequency of the first class, relative to the frequency with which the hypothesis is true, is calculable, and therefore controllable simply from the specification of the null hypothesis. The frequency of the second kind must depend not only on the frequency with which rival hypotheses are in fact true, but also greatly on how closely they resemble the null hypothesis. Such errors are therefore incalculable both in frequency and in magnitude merely from the specification of the null hypothesis, and would never have come into consideration in the theory only of tests of significance, had the logic of such tests not been confused with that of acceptance procedures.

It may be added that in the theory of estimation we consider a continuum of hypotheses each eligible as null hypothesis, and it is the aggregate of frequencies calculated from each possibility in turn as true—including frequencies of error, therefore only of the "first kind", without any assumptions of knowledge *a priori*—which supply the likelihood function, fiducial limits, and other indications of the amount of information available. The introduction of allusions to errors of the second kind in such arguments is entirely formal and ineffectual.

The fashion of speaking of a null hypothesis as "accepted when false", whenever a test of significance gives us no strong reason for rejecting it, and when in fact it *is* in some way imperfect, shows real ignorance of the research workers' attitude, by suggesting that in such a case he has come to an irreversible decision.

The worker's real attitude in such a case might be, according to the circumstances:

(*a*) "The possible deviation from truth of my working hypothesis, to examine which the test is appropriate, seems not to be of sufficient magnitude to warrant any immediate modification." Or it might be:

(*b*) "The deviation is in the direction expected for certain influences which seemed to me not improbable, and to this extent my suspicion has been confirmed; but the body of data available so far is not by itself sufficient to demonstrate their reality."

These examples show how badly the word "error" is used in describing such a situation. Moreover, it is a fallacy, so well known as to be a *standard* example, to conclude from a test of significance that the null hypothesis is thereby established; at most it may be said to be confirmed or strengthened.

In an acceptance procedure, on the other hand, acceptance is irreversible, whether the evidence for it was strong or weak. It is the result of applying mechanically rules laid down in advance;

no *thought* is given to the particular case, and the tester's state of mind, or his capacity for *learning*, is inoperative.

By contrast, the conclusions drawn by a scientific worker from a test of significance are *provisional*, and involve an intelligent attempt to *understand* the experimental situation.

4. "*Inductive Behaviour*"

The erroneous insistence on the formula of "repeated sampling from the same population" and the misplaced emphasis on "errors of the second kind" seem both clearly enough to flow from the notion that the process by which experimenters learn from their experiments might be equated to some equivalent acceptance procedure. The same confusion evidently takes part in the curious preference expressed by J. Neyman for the phrase "inductive behaviour" to replace what he regards as the mistaken phrase "inductive reasoning".

Logicians, in introducing the terms "inductive reasoning" and "inductive inference" evidently imply that they are speaking of processes of the mind falling to some extent outside those of which a full account can be given in terms of the traditional deductive reasoning of formal logic. Deductive reasoning in particular supplies no essentially new knowledge, but merely reveals or unfolds the implications of the axiomatic basis adopted. Ideally, perhaps, it should be carried out mechanically. It is the function of inductive reasoning to be used, in conjunction with observational data, to add new elements to our theoretical knowledge. That such a process existed, and was possible to normal minds, has been understood for centuries; it is only with the recent development of statistical science that an analytic account can now be given, about as satisfying and complete, at least, as that given traditionally of the deductive processes.

When, therefore, Neyman denies the existence of inductive reasoning he is merely expressing a verbal preference. For him "reasoning" means what "deductive reasoning" means to others. He does not tell us what in his vocabulary stands for inductive reasoning, for he does not clearly understand what that is. What he tells us to call "inductive behaviour" is merely the practice of making some assertion of the form

$$T < \theta$$

in some circumstances, and refraining from this assertion in others. This is evidently an effort to assimilate a test of significance to an acceptance procedure. From a test of significance, however, we learn more than that the body of data at our disposal would have passed an acceptance test at some particular level; we may learn, if we wish to, and it is to this that we usually pay attention, at what level it would have been doubtful; doing this we have a genuine measure of the confidence with which any particular opinion may be held, in view of our particular data. From a strictly realistic viewpoint we have no expectation of an unending sequence of similar bodies of data, to each of which a mechanical "yes or no" response is to be given. What we look forward to in science is further data, probably of a somewhat different kind, which may confirm or elaborate the conclusions we have drawn; but perhaps of the same kind, which may then be added to what we have already, to form an enlarged basis for induction.

Neyman reinforces his choice of language by arguments much less defensible. He seems to claim that the statement (*a*) "θ has a probability of 5 per cent. of exceeding T" is a different statement from (*b*) "T has a probability of 5 per cent. of falling short of θ". Since language is meant to be used I believe it is essential that such statements, whether expressed in words or symbols, should be recognized as equivalent, even when θ is a parameter, defined as an objective character of the real world, entering into the specification of our hypothetical population, whilst T is directly calculable from the observations. To prevent the kind of confusion that Neyman has introduced we may point out that both statements are statements of the relationship in which T, or θ, stands to the other. Also, since *probability* is specified, the statements have meaning only in relation to a sufficiently well-defined population of pairs of these values. The statements do not imply that in this population of pairs of values either T or θ is constant, but also they do not exclude the possibility that one should be constant, and that variability should be confined to the other. Reference to the mode of calculating our limits in an ordinary test of significance will generally establish that in these calculations the parameter θ has been treated provisionally as constant, and variations calculated of T for given θ. The possible variation of θ is left arbitrary, and is irrelevant to the calculations, much as is the distribution of the independent variate in the regression problem.

A complementary doctrine of Neyman violating equally the principles of deductive logic is to accept a general symbolical statement such as

$$Pr\{(\bar{x} - ts) < \mu < (\bar{x} + ts)\} = \alpha\,,$$

as rigorously demonstrated, and yet, when numerical values are available for the statistics $\bar{x}$ and s, so that on substitution of these and use of the 5 per cent. value of t, the statement would read

$$Pr\{92{\cdot}99 < \mu < 93{\cdot}01\} = 95 \text{ per cent.},$$

to *deny* to this *numerical* statement any validity. This evidently is to deny the syllogistic process of making a substitution in the major premise of terms which the minor premise establishes as equivalent. By this, which is surely a desperate measure, Neyman supports the assertion that if μ stand for some objective constant of nature, or property of the real world, such as the distance of the sun, its probability of lying between any named numerical limits is necessarily either 0 or 1, and we cannot know which, unless the true distance is known to us. The paradox is rather childish, for it requires that we should wilfully misinterpret the probability statement so as to pretend that the population to which it refers is not defined by our observations and their precision, but is absolutely independent of them. As this is certainly not what any astronomer means, and is not in accordance with the origin of the statement he makes, it seems rather like an acknowledgement of bankruptcy to pretend that it is.

Finally let me add some notes on what appear to me to be distinctive requirements of valid inductive inference.

5. *Requirements of Inductive Inferences*

(*a*) Since some inductive inferences are expressed in terms of *probability* (fiducial probability) the first requirement is a clear understanding that probability statements always have reference to some sufficiently defined population, and never to individuals, save as typical members of such a population. This understanding is needed for deductive inferences also, when statements of probability are made.

(*b*) A very important feature of inductive inference, unknown in the field of deductive inference, is the framing of the hypothesis in terms of which the data are to be interpreted. This hypothesis must fulfill several requirements: (i) it must be in accordance with the facts of nature as so far known; (ii) it must specify the frequency distribution of all observational facts included in the data, so that the data as a whole may be taken as a typical sample; (iii) it must incorporate as parameters all constants of nature which it is intended to estimate, in addition possibly to special, or *ad hoc*, parameters; (iv) it must not be contradicted, *in any way judged relevant*, by the data in hand. If it satisfies these conditions it is therefore a scientific construct of a fairly elaborate type. It is by no means obvious that different persons should not put forward different successful hypotheses, among which the data can supply little or no discrimination. The hypothesis is sometimes called a model, but I should suggest that the word model should only be used for aspects of the hypothesis between which the data cannot discriminate. As an act of construction the hypothesis is not altogether impersonal, for the scientist's personal capacity for theorizing comes into it; moreover, the criteria by which it is approved require a certain honesty, or integrity, in their application.

(*c*) In one respect inductive reasoning is more strict than is deductive reasoning, since in the latter any item of the data may be ignored, and valid inferences may be drawn from the rest; i.e. from any selected sub-set of the set of axioms used, whereas in inductive inference the whole of the data must be taken into account. This seems to be very difficult to be understood by workers trained in deductive methods only, though more easily understood by statisticians. The political principle that anything can be proved by statistics arises from the practice of presenting only a selected sub-set of the data available.

In some early results of my own I rely on the datum "There is no knowledge of probabilities *a priori*". They would not certainly have been legitimate without this datum, but they have been mistakenly described as a kind of greatest common factor of the inferences which could be drawn for different possible data giving probabilities *a priori*.

It is revealing that this logical distinction was overlooked by Neyman and Pearson, in 1933, in one of their earliest papers after they had learnt of the possibility of inferring fiducial limits, the argument for which I had set out in a paper on *inverse probability* in the *Proceedings of the Cambridge Philosophical Society*, 1930. It is particularly instructive that although in that paper I speak of "learning by experience", of "inductive processes", and of "the probability of causes", much as others had done since the eighteenth century, these authors read into my work "rules of behaviour", which indeed I had not mentioned at all. Both misapprehensions become intelligible if we realise that the authors had no idea of a test of significance as a means of learning, but conceived it only under the form of an acceptance procedure. The passage is as follows:

"In a recent paper [(Neyman & Pearson, 1933*b*)] we have discussed certain general principles underlying the determination of the most efficient tests of statistical hypotheses, but the method of approach did not involve any detailed consideration of the question of *a priori* probability. We propose now to consider more fully the bearing of the earlier results on this question and in particular to discuss what statements of value to the statistician in reaching his final judgement can be made from an analysis of observed data, which would not be modified by any change in the probabilities *a priori*. In dealing with the problem of statistical estimation, R. A. Fisher has shown how, under certain conditions, what may be described as *rules of behaviour* can be employed which will lead to results independent of these probabilities; in this connection he has discussed the important conception of what he terms fiducial limits.[8, 9]. But the testing of statistical hypotheses cannot be treated as a problem in estimation, and it is necessary to discuss afresh in what sense tests can be employed which are independent of *a priori* laws."

There seems here an entirely genuine inability to conceive that when new data are added in an inductive problem, previously correct conclusions are no longer correct. Or, in this case that the conclusions proper to the absence of knowledge of probabilities *a priori* would be wrong for almost any set of such probabilities, and could in no sense be a common term in the proper inferences from all such sets.

(*d*) Variety of logical form.

A fourth feature which has emerged in the study of inductive inference is that data of apparently the same logical form, though with different mathematical specification, give rise to inferences not always of the same logical form.

For example, when in 1930 I introduced the notions of the fiducial distribution and fiducial limits I did so with the example of the sampling distribution of the estimated correlation coefficient r for various values of the true correlation ρ. The distribution of r is continuous between the limits -1 and $+1$, and for any value of P there is a value of r, which may be called $r_P(\rho)$, such that r exceeds it with frequency $1-P$, and falls short of it with frequency P. These functions of ρ increase monotonically from -1 to $+1$ as ρ passes from -1 to $+1$. Consequently, corresponding with any observed value r, there is a value of ρ, which may be denoted as $\rho_{1-P}(r)$ such that for this value of ρ the observed value will fall short of r with frequency P and exceed it with frequency $1-P$. In fact if P is expressed as an explicit function

$$P = F_N(r, \rho)$$

such that the distribution of r for given ρ is given by the frequency element

$$\frac{\partial F}{\partial r}\,dr,$$

then the distribution

$$-\frac{\partial F}{\partial \rho}\,d\rho$$

will be the fiducial distribution of ρ for given r, in the sense that the frequency of exceeding any chosen value of ρ is the frequency, for that value of ρ, of r being less than the value observed. The quantiles of this distribution thus give the fiducial limits of ρ at any chosen level of significance.

Had I taken a discontinuous variate, such as the number of successes observed out of N trials, and sought in terms of the observations to obtain a fiducial distribution for the true probability, (say x), it would certainly have been possible to find a value of x such that the probability of the number of successes observed, or any higher number was, let us say 5 per cent., so that smaller

values of x could be rejected at least at the 5 per cent. level of significance; but this gives only an inequality statement for the probability that x is less than any given value. Neyman seems to ignore this distinction, and to speak in both cases of confidence limits. Logically, however, the form of inference admissible is totally distinct.

Equally, statements of fiducial probability in continuous cases are only proper if the whole of the information is utilized, as it is by the use of sufficient estimates, whereas for any test of significance, however low in power, it may well be possible to point to the limits outside which parametric values are significantly contradicted by the data at a given level of significance. These also should be regarded as giving only rough statements for the fiducial probability.

There are other cases in the theory of estimation in which rather similar data yield information of remarkably different kinds. Consider, for example, the case in which x and y are two observables distributed in normal distributions with unit variance in each case, and independently, about hypothetical means ξ and η. No situation could be simpler. Suppose, however, that the data contain a functional relationship connecting ξ and η. Then different cases arise from different functional forms:

(i) If there is a simple linear connection between ξ and η, so that (ξ, η) represents a point on a given straight line, then the foot of the perpendicular from the observation point (x, y) is a sufficient estimate, and the fiducial distribution of (ξ, η) on the given line will be a normal distribution with unit variance about this estimate. All possible observations on the same perpendicular are equivalent.

(ii) If the given locus of (ξ, η) is a circle, there is no sufficient estimate; the distance of (x, y) from the centre of the given circle is, however, an ancillary statistic, which together with the maximum likelihood estimate makes the estimation exhaustive. For each possible distance an appropriately oriented fiducial distribution on the circle may be specified.

(iii) In general there is a well defined likelihood function, and therefore an estimated point of maximum likelihood. It is not obvious that any general substitute can be found for the ancillary statistic, save in an asymptotic sense, or that any statement of fiducial probability is possible in general. Thus three logically distinct types of inference arise from simple changes in the mathematical specification of the problem.

(*e*) Finally, in inductive inference we introduce no cost functions for faulty judgements, for it is recognized in scientific research that the attainment of, or failure to attain to, a particular scientific advance this year rather than later, has consequences, both to the research programme, and to advantageous applications of scientific knowledge, which cannot be foreseen. In fact, scientific research is not geared to maximize the profits of any particular organization, but is rather an attempt to improve *public* knowledge undertaken as an act of faith to the effect that, as more becomes known, or more surely known, the intelligent pursuit of a great variety of aims, by a great variety of men, and groups of men, will be facilitated. We make no attempt to evaluate these consequences, and do not assume that they are capable of evaluation in any sort of currency.

When decision is needed it *is* the business of inductive inference to evaluate the *nature* and *extent* of the uncertainty with which the decision is encumbered. Decision itself must properly be referred to a set of motives, the strength or weakness of which should have had no influence whatever on any estimate of probability. We aim, in fact, at methods of inference which should be equally convincing to all rational minds, irrespective of any intentions they may have in utilizing the knowledge inferred.

We have the duty of formulating, of summarising, and of communicating our conclusions, in intelligible form, in recognition of the right of *other* free minds to utilize them in making *their own* decisions.

References

BARNARD, G. A. (1945), "A new test for 2 × 2 tables", *Nature*, **156**, No. 3954, 177.
—— (1946), "Sequential tests in industrial statistics", *J.R. Statist. Soc., Supp* . **8**, 1–21.
—— (1947*a*), "Significance tests for 2 × 2 tables", *Biometrika*, **34**, 123–138.
—— (1947*b*), "The meaning of a significance level", *Biometrika*, **34**, 179–182.
—— (1947*c*), Review: Sequential Analysis. By Abraham Wald, *J. Amer. Stat. Ass.*, **42**, 658.
—— (1949), "Statistical inference", *J. R. Statist. Soc., B*, **11**, 115–139.
COCHRAN, W. G., & COX, G. M. (1950), *Experimental Designs*. New York: Wiley. London: Chapman & Hall.

DAVIES, O. L. (ed.) (1954), *The Design and Analysis of Industrial Experiments*. London & Edinburgh: Oliver & Boyd.
FINNEY, D. J. (1952), *Statistical Method in Biological Assay*. London: Griffin.
FISHER, R. A. (1922), "The goodness of fit of regression formulae, and the distribution of regression coefficients", *J.R. Statist. Soc.*, **85**, 597–612.
—— (1930), "Inverse probability", *Proc. Camb. Phil. Soc.*, **26**, 528–535.
—— (1933), "The concepts of inverse probability of fiducial probability referring to unknown parameters", *Proc. Roy. Soc., A*, **139**, 343–348.
—— (1934), *Statistical Methods for Research Workers*. (5th ed. and later.) London & Edinburgh: Oliver & Boyd.
—— (1941), "The interpretation of experimental fourfold tables", *Science*, **94**, No. 2435, 210–211.
—— (1945), "A new test for 2 × 2 tables", *Nature*, **156**, No. 3961, 388.
GOULDEN, C. H. (1939 and 1952), *Methods of Statistical Analysis*. New York: Wiley. London: Chapman & Hall.
NEYMAN, J. (1938), "L'estimation statistique traité comme un problème classique de probabilité", *Actualités Scientifiques et Industrielles*, No. 739, 25–57.
—— & PEARSON, E. S. (1933*a*), "The testing of statistical hypotheses in relation to probabilities *a priori*", *Proc. Camb. Phil. Soc.*, **29**, 492–510.
—— (1933*b*), "On the problem of the most efficient tests of statistical hypotheses", *Phil. Trans. Roy. Soc., A*, **231**, 289–337.
PEARSON, E. S. (1947), "The choice of statistical tests illustrated on the interpretation of data classed in a 2 × 2 table", *Biometrika*, **34**, 139–167.
"STUDENT" (1908), The probable error of a mean, *Biometrika*, **6**, 1–25.
VENN, J. A. 1876), *The Logic of Chance* (2nd ed.). London: Macmillan.
WALD, A. (1950), *Statistical Decision Functions*. New York: Wiley. London: Chapman & Hall.
WILSON, E. B. (1941), "The controlled experiment and the fourfold table", *Science*, **93**, No. 2424, 557–560.
—— (1942*a*), "On contingency tables", *Proc. Nat. Acad. Sci.*, **28**, No. 3, 94–100.
WORCESTER, J. (1942*b*), "Contingency tables", *Proc. Nat. Acad. Sci.*, **28**, No. 9, 378–384.
YATES, F. (1949), *Sampling Methods for Censuses and Surveys*. London: Griffin.