The Fisher/Pearson Chi-Squared Controversy: A Turning Point for Inductive Inference*

by DAVIS BAIRD

R. A. Fisher introduced 'degrees of freedom' to resolve the Chi-Squared Controversy. When degrees of freedom are accounted for in test construction *significant outcomes* are evidence that the hypothesis under test is inadequate, not simply false. At least two criteria measure the adequacy of an hypothesis: faithfulness to observations and informativeness. Degrees of freedom measure the informativeness of an hypothesis. An uninformative hypothesis which is extremely faithful to observations can be less adequate than an informative hypothesis which is not as faithful to observations. By accounting for degrees of freedom significance tests answer to this rule of thumb.

I write about a small incident in the history of statistics. In 1915 G. U. Yule and M. Greenwood discovered that two widely used statistical tests yield contradictory inferences regarding independence in a two-by-two contingency table. By the late teens Yule, through a series of experiments, localised the problem to what are now called 'degrees of freedom'. R. A. Fisher published a mathematical analysis in 1922; his analysis concluded that the Chi-Squared test must be slightly changed. Karl Pearson, the inventor of the Chi-Squared test, objected to Fisher's solution. Pearson's objections fell on deaf ears. Chi-Squared is now applied as Fisher argued it ought to be.

The Chi-Squared Controversy seems trivial: a contradiction is discovered; the problem is localised; a mathematical analysis resolves the issue; no large scale overhaul in practice is necessary and no substantial disagreement ensues. However, the Chi-Squared Controversy marks a turning point

Received 28 September 1981

* I would like to thank Professor Ian Hacking who first suggested that I investigate the Fisher/Pearson Chi-Squared Controversy and who criticised many earlier drafts of this paper.

in inductive inference. Prior to this Controversy it was possible to interpret statistical tests as a formal means of showing statistical hypotheses to be false (or likely false) in the light of evidence. The introduction of degrees of freedom blocks this interpretation. Now statistical tests are used to show hypotheses inadequate in the light of evidence. A false hypothesis need not be inadequate.

What is at issue may be put into relief by focusing on a rule of inductive inference commonly called the Consequence Condition. This rule appears in many forms depending on the specific approach to inductive inference taken. Here is one version:

> (*CC*) If hypothesis h implies hypothesis i and evidence e warrants the rejection of i, then e warrants the rejection of h.

Carl Hempel calls such a rule a condition of adequacy for any theory of confirmation (Hempel [1965], p. 31). Hempel is not alone: Karl Popper ([1959], p. 76), Ian Hacking ([1976], p. 33), L. J. Savage ([1970], p. 32) and R. A. Fisher ([1959], p. 21) all adopt some form of the Consequence Condition. It is tempting to call the Consequence Condition a first principle of inductive inference. Prior to the Chi-Squared Controversy, the Chi-Squared test did not violate the Consequence Condition; after Fisher's introduction of degrees of freedom, Chi-Squared violated the Consequence Condition. And so matters stand today.

1 THE CHI-SQUARED TEST

The Chi-Squared test measures the goodness of fit of a statistical hypothesis. For example, the hypothesis that a die is fair does not well fit the following data:

ace	2	3	4	5	6
1	1	1	1	1	20

Chi-Squared provides one plausible measure of how well an hypothesis fits frequency data.

The textbook application of Chi-Squared requires data which may be sorted into n disjoint and mutually exhaustive possible outcomes or 'cells', *i.e.*, a multinomial chance set-up. The hypothesis tested stipulates, or at least partially constrains, the chance of any single trial of the chance set-up being in the different cells. If the hypothesis stipulates exact cell probabilities for each cell, it is called 'simple'. Hypotheses which only partially constrain the cell probabilities are called 'composite'.

A simple hypothesis stipulates cell probabilities, p_i, for each cell, i. Given S independent trials of the chance set-up the number of outcomes *expected* on the basis of the hypothesis in each cell i is Sp_i. The Chi-Squared statistic,

χ^2, sums the squared differences between the number of outcomes expected in cell i, Sp_i, and the number which did in fact occur in cell i, s_i:

Cell	1	2	3	$\ldots n$	
Observed	s_1	s_2	s_3	$\ldots s_n$	$\sum_i s_i = S$
Expected	Sp_1	Sp_2	Sp_3	$\ldots Sp_n$	$\sum p_i = 1$

$$\chi_0^2 = \sum_i (S_i - Sp_i)^2 / Sp_i$$

Given the hypothesis, we can determine the limiting probability as the number of observations, S, grows without bound of a given Chi-Squared value χ_0^2 or higher:

$$\mathrm{P}(\chi^2 \geqslant \chi_0^2) = \int_{\chi_0^2}^{\infty} \left(1/2^{1/2(n)} \Gamma\left(\frac{n}{2}\right)\right) e^{-1/2\chi^2} \chi^{2^{1/2(n-2)}} d\chi^2$$

This limiting probability provides an excellent approximation for values of S given by more than 10 observations per cell. The exact density function used in any specific case is given by the number of cells $(n+1)$. $\mathrm{P}(\chi^2 \geqslant \chi_0^2)$ is the probability of observations as discrepant from expectations, or worse than the discrepancy observed, χ_0^2. If this probability, called the *level of significance*, is very small, less than, for example, 0.05 or 0.01, we reject the hypothesis. The family of Chi-Squared density functions is well tabled. Each specific set of probabilities, $\mathrm{P}(\chi^2 \geqslant \chi_0^2)$, is given by $(n+1)$, the number of cells. Tables in hand, Chi-Squared is easily applied.

Many justifications for this procedure have been published. Karl Pearson connected the procedure with his theory of multivariate normal correlation (K. Pearson [1900]). Within his positivist philosophy, correlation replaced causation. Thus for Pearson, Chi-Squared measured causal connectedness. Much of Pearson's philosophy is abandoned now, but contemporary justifications may be found. Chi-Squared is the limit of multinomial probabilities in the same way that the Normal density is the limit of binomial probabilities. Chi-Squared is also the limiting case of likelihood ratio tests (Mood *et al.* [1974], p. 444). Indeed most foundational approaches embrace Chi-Squared. Beyond theoretical justification, Chi-Squared has seen more use than any other statistical test beside possibly 'Student's' *t*-test.

When the hypothesis under test is composite the mechanics of testing get more complicated. Such hypotheses do not provide enough information to compute expected cell frequencies. Fisher introduced degrees of freedom to cope with the difficulties present in testing composite hypotheses. Pearson objected. I focus on some of these arguments below.

2 YULE AND GREENWOOD'S 1915 PAPER

The Chi-Squared Controversy begins with Yule and Greenwood's 1915 paper on the efficacy of cholera and typhoid vaccination. To test efficacy, Yule and Greenwood use data represented in the following 'contingency table':

	attacked	not	total
vaccinated	a	b	$a+b$
not	c	d	$c+d$
total	$a+c$	$b+d$	$a+b+c+d=N$

Yule and Greenwood observe that if vaccination is statistically independent of attack, then the vaccine is not helpful. They test the hypothesis of statistical independence with the Chi-Squared test. The contingency table has four cells, so they use probabilities taken from the Chi-Squared table for four cells ($n+1=4$).

Yule and Greenwood note that the conditions of their test do not meet all of the theoretical assumptions Pearson used to construct the test. The hypothesis of statistical independence does not stipulate exact cell probabilities; it merely constrains the probabilities such that P(attacked & vaccinated) = P(attacked) P(vaccinated). They write:

> It must be noticed that the application of this test to innoculation data is based on an assumption. We do not in fact know the true values of α and β (chance of being innoculated and chance of escaping disease) and must replace them by the observed ratios of the number of innoculated persons to the total frequency and the number of cases of disease (or deaths) to the total. This is not strictly correct, from the point of view of general theory, but when we are dealing with such distributions as those actually in question, there is, perhaps no more impropriety in making the assumption than in following the same course when we compute errors of simple sampling in the ordinary way (Yule and Greenwood [1915], p. 118).

It is necessary to have exact probabilities for attack, and for vaccination, in order to compute expected cell frequencies—and thereby to compute the value of Chi-Squared, χ_0^2. Since these probabilities are not stipulated by the hypothesis, Yule and Greenwood suggest that they should be approximated by observed frequencies of attack and of vaccination. Thus, given the data:

	attacked	not	marginal totals
vaccinated	a	b	$a+b$
not	c	d	$c+d$
marginal totals	$a+c$	$b+d$	$a+b+c+d=N$

P(attacked) would be approximated by the value $(a+c)/N$; P(vaccinated) would be approximated by $(a+b)/N$. Their 'expected' probabilities are determined from the observed marginal totals of the evidence used to test the hypothesis.

This approximation procedure fixes the marginal totals of 'expected' cell frequencies to the observed marginal totals. It is possible to compute exact expected cell frequencies, a', b', c' and d' by assuming: 1) $a'+b' = a+b$—thus approximating the value of P(vaccinated); 2) $a'+c' = a+c$—thus approximating the value of P(attacked); and 3) statistical independence: P(attacked & vaccinated) $= a'/N = ((a'+b')/N)((a'+c'/N) = ((a+b)/N)$ $((a+c)/N) = P$(attacked) P(vaccinated). Thus, while it is impossible to compute exact expected cell frequencies on the hypothesis of statistical independence alone, exact expected cell frequencies can be computed with this additional method of approximating P(attacked) and P(vaccinated). This additional assumption is expedient and, initially, quite plausible.

Matters could have perhaps rested here if Yule and Greenwood had not encountered an unexpected problem:

> In our subsequent discussion we shall frequently compare the two ratio's innoculated attacked/all innoculated with uninnoculated attacked/all uninnoculated which we may denote p_1 and p_2. It may therefore be asked why we should not adopt as our criterion of significance the ratio of $p_1 - p_2$ to its standard error, counting as significant all differences greater than some assigned multiple of the standard error. It will be found that if this plan is adopted deviations which judged by the χ^2 test are not improbable are much less likely to occur as the result of random sampling (*ibid.*).

They further observe that probabilities from the Chi-Squared table for two cells better agree with 'standard error test' probabilities than do probabilities from the Chi-Squared table for four cells.

They suggest an explanation: the 'expected' marginal totals were artificially constrained to equal those observed. This restricts the ways in which observed frequencies might deviate from 'expected' frequencies; the 'degrees of freedom' are reduced. 'Degrees of freedom', however, is a concept yet to be clearly understood.

Yule and Greenwood proceeded to use the more cautious Chi-Squared table for four cells. However there is a significant discrepancy. Different attitudes towards rejection follow from the contradicting tests:

	attacked	not	totals
vaccinated	409	3	412
not	174	8	182
totals	583	11	594

$\chi_0^2 = 9.34$ [when expected frequencies are computed as indicated above]

$P(\chi^2 \geqslant 9.34) = 0.02442$ [using the table for four cells]

$P(\chi^2 \geqslant 9.34) = 0.00217$ [using the table for two cells]

3 FISHER'S ARGUMENT

In 1922 R. A. Fisher published an exact solution (Fisher [1922]). The cell frequencies from which χ_0^2 is computed are not, argued Fisher, simply approximations. These frequencies are functions of the sample data. This functional relationship constrains the manner in which observed values may deviate from 'expected' values. On Fisher's view, it is not merely the number of cells which determines the exact χ^2 density curve—and hence table for finding probabilities—but all constraints on the deviations. The number of cells provides one constraint; fixed marginal totals provide further constraints.

Fisher's argument is easily understood upon a more detailed examination of the uncontroversial case. Here one is concerned to determine the density function for

$$\chi^2 = \sum_i \frac{(s_i - Sp_i)^2}{Sp_i}$$

Writing e_i for $s_i - Sp_i$, we have

$$\textstyle\sum_i (e_i^2/Sp_i)$$

The e_i's, or cell deviations, may be jointly represented in an $n-1$ dimensional space; $n-1$ of the e_i's may take any value; then nth is constrained:

$$1)\ \textstyle\sum_i e_i = \sum_i (s_i - Sp_i) = \sum_i s_i - \sum_i Sp_i = S - S\sum_i p_i = S - S = 0$$

By means of geometrical proof, the density function is shown to be

$$f(\chi^2) = \frac{1}{2^n\Gamma\left(\frac{n}{2}\right)} e^{-1/2\chi^2}\chi^{2^{1/2(n-2)}}$$

where n is the number of dimensions necessary to represent the e_i's, and not simply the number of cells minus one. This construction is displayed in greater detail in Kendall and Stuart's *The Advanced Theory of Statistics* (Kendall and Stuart [1977], vol. 1, p. 271). One of Fisher's greatest contributions to statistics is just this method of geometrical representation and proof.

In standard cases, $n-1$ dimensions are required to represent the e_i's. However, if the expected cell frequencies are not independent of the observed frequencies fewer dimensions are required. Fisher's 1922 proof consists in pointing out that when the expected frequencies are estimated from the observed frequencies, on the basis of fixing the marginal totals, for example, this independence is violated.

Take Yule and Greenwood's problem, with the following data:

	B	$\bar{B}$	total
A	a	b	$a+b$
$\bar{A}$	c	d	$c+d$
total	$a+c$	$b+d$	$a+b+c+d$

In order to apply Chi-Squared they construct a second contingency table consistent with the hypothesis of independence, and with the same marginal totals as those above. The e_i's are constrained beyond 1) $\sum_i e_i = 0$; 2) $e_1+e_2 = (a-a')+(b-b') = (a+b)-(a'+b') = (a+b)-(a+b) = 0$. Similarly, 3) $e_3+e_4 = 0$; 4) $e_1+e_3 = 0$; 5) $e_2+e_4 = 0$. Three of these five equations are independent. Four dimensions are necessary to represent all possible values of the vector (e_1, e_2, e_3, e_4). However, once these three independent linear equations are imposed, only one dimension is necessary. Fisher calls the number of dimensions necessary to represent the deviations between an hypothesis and evidence the 'degrees of freedom' of the hypothesis. Thus, the density curve for χ^2 for a two-by-two contingency table with estimated expected frequencies on the basis of fixed marginal totals is the same as that for a two-celled multinomial with no estimation. The hypothesis of statistical independence has one degree of freedom, not three.

4 PEARSON'S REPLY

Fisher's proof was a direct attack on one of Pearson's greatest achievements: the Chi-Squared test. It was clear that not only must the table for two cells given no estimation be used in the contingency table case, but a different table must be used in any case where exact probabilities are estimated from an observed sample. Pearson invented the practice of estimating exact curves of a given Pearsonian type by the method of sample moments. All of the important uses of the Chi-Squared test involved estimated frequencies. Fisher argues that the Chi-Squared test must be changed, not solely in the case of contingency tables, but quite generally. According to Fisher, Pearson's 1900 paper makes a fundamental mistake.

Pearson replied swiftly with great hostility in the 1922 issue of *Biometrika*:

> The above re-description of what seems to me very elementary considerations would be unnecessary had not a recent writer in the *Journal of the Royal Statistical Society* appeared to have wholly ignored them. He considers that I have made serious blunders in not linking my degrees of freedom by the number of moments I have taken; . . . I hold that such a view is entirely erroneous and that the writer has done no service to the science of statistics by giving it broad cast circulation in the pages of the *Journal of the Royal Statistical Society*.

Pearson concluded:

> I trust my critic will pardon me for comparing him with Don Quixote tilting at the windmill; he must either destroy himself or the whole of the theory of probable errors, . . . (Pearson [1922], p. 187).

Besides rhetoric, Pearson presents two main points: 1) When the hypothesis of interest is statistical independence, the true χ_0^2 value cannot be determined; an approximate value may be obtained by holding the marginals fixed. 2) The mean-squared contingency, $\phi^2 = \chi^2/N$, does not result from sampling; it is a means for determining correlation. Thus it is incorrect to treat data as a sample and thereby to argue that methods of estimation have probabilistic properties.

In his 1900 paper presenting Chi-Squared for the first time, Pearson defines Chi-Squared in terms of exact hypothesised cell probabilities. Most hypotheses do not stipulate exact cell probabilities. Consequently there is no way to determine the true value of χ_0^2. All is not lost. An approximate value for Chi-Squared may be found. The approximate value is found by using observed frequency data and the constraints imposed by the hypothesis. The method for dealing with the hypothesis of independence is one example. Thus Pearson writes:

> What we actually do is replace the accurate value of χ^2, which is unknown to us, and cannot be found, by an approximate value, and we do this with precisely the same justification as the astronomer claims, when he calculates his probable error on his observations, and not on the mean square error of an infinite population of errors which is unknown to him (*ibid.*).

Of course, the probability associated with the approximate Chi-Squared value no longer represents the 'true' probability of a random sample deviating from expected cell frequencies. Pearson, however, never announces this fact.

Pearson's failure to be clear on what the Chi-Squared probability represents is not overly surprising. For Pearson Chi-Squared is a *measure of fit*; the probabilities associated with the test do not represent anything like 'the frequency of incorrectly rejecting the hypothesis'. They provide a convenient way to describe the fit between hypothesis and data. Indeed probabilities need be introduced so that uses of Chi-Squared with four cells, for example, can be objectively compared with uses with fourteen cells. Furthermore, this is consistent with Pearson's subjective interpretation of probability.

Fisher's proof is astonishing. A true χ_0^2 *can* be obtained even when expected cell frequencies must be estimated. The argument rests on Fisher's analysis of estimation which, at the time, was revolutionary. Pearson did not accept Fisher's analysis of estimation; he retained his different notion of approximation: the true value of χ_0^2 cannot be obtained. Pearson's second argument buttresses this point.

To understand Pearson's second argument a short aside is necessary.

Pearson extended his original 1900 Chi-Squared test to cope with contingency table data in his [1904]. In this paper the Chi-Squared statistic is part of an argument purporting to show how discontinuous contingency table data can be related to continuous, Normally Correlated, random variables. Let $\phi^2 = \chi^2/N$, where χ^2 is computed as above for contingency table data and N is the total number of trials. Pearson ([1904], p. 448) argues that

$$r_{xy} = (\phi^2/(1+\phi^2))^{1/2}$$

where r_{xy} is the correlation between x and y (attack rates and vaccination rates). This result is important because, according to Pearson, correlation replaces causation. Thus if his argument is correct, Pearson has shown how to measure the causal relationship between, for example, vaccination and cholera attack.

The use of χ^2 in this argument is purely formal. It is not to be understood as a random variable for which probabilities may be determined. Thus Fisher's argument, in claiming to determine a probability for χ^2 when computed from a contingency table, misunderstands the nature of a contingency χ_0^2 value. Consequently, whatever means Fisher used to find a probability density function for χ^2 computed from a contingency table must be wrong.

5 ASSESSMENT

Pearson's counter to Fisher is invalid. In the first place Fisher demonstrates correctly just what Pearson says is impossible: when expected cell frequencies are estimated on the basis of fixed marginal totals the Chi-Squared table for two cells correctly represents the probabilities of the resulting Chi-Squared statistic. A 'true' Chi-Squared value and hence probability may be found. Pearson's second point about ϕ^2 and Fisher's incorrect interpretation of contingency table χ_0^2 values fails because his 1904 argument is invalid. It is not possible in general to reduce measures of association in a contingency table to Pearson's measure of Normal Correlation r_{xy}.

On the other hand, Fisher's argument is also insufficient to establish his claim. Fisher proves that $\chi_0^2 = \sum_i((s_i - f(s_i))^2/f(s_i))$, where $f(s_i)$ is the expected cell frequency computed from the contingency table sample, has the same density function as that for the Chi-Squared statistic for two cells and no estimation. Fisher fails to show that $\chi_0^2 = \sum_i((s_i - f(s_i))^2/f(s_i))$ is a plausible measure of goodness of fit between the hypothesis of independence and contingency table observations. I do not claim that the Chi-Squared statistic with expected cell frequencies functionally related to observed cell frequencies is *not* a plausible measure of fit. I claim only that Fisher has not shown it to be so.

It is not obvious that any *a priori* argument exists to settle the Chi-Squared Controversy. What is a plausible measure of fit between hypothesis and evidence depends on what criteria are important for fit. Neither Pearson nor Fisher discuss this crucial question of what constitutes a good fit. I do not claim to know what ought to constitute criteria of good fit. I do make two claims: 1) If, by fit, we want to measure simply the difference between predicted and observed values—the 'closeness to the truth' of the hypothesis—then Pearson's method of Chi-Squared testing is better than Fisher's. 2) If, by fit, we want to measure both the closeness to the truth of the hypothesis and the informativeness of the hypothesis, then Fisher's method is better than Pearson's.

6 GOODNESS OF FIT AND CLOSENESS TO TRUTH

The composite hypothesis, *SI*, of statistical independence, and composite hypotheses generally, may be understood as a disjunction of (infinitely) many simple hypotheses. If the composite hypothesis, *SI*, is true, then there are some exact probabilities, *P*(attacked) and *P*(vaccinated) which satisfy *P*(attacked) *P*(vaccinated) = *P*(attacked & vaccinated); if it is false, then there are no such probabilities. We can notate the composite hypothesis *SI*: (h_i v h_2 . . . v h_ω v . . .). Each h_i represents one set of exact probabilities satisfying the constraint of statistical independence. If *SI* is true then one (and only one) member, h_t, of this disjunction is true; if *SI* is false, then all the h_i's are false.

This observation suggests one manner for dealing with composite hypotheses. If by 'fit' we mean something like 'closeness to the truth' of an hypothesis, then rejecting an hypothesis because of bad fit would be accomplished by rejecting each disjunct h_i individually on account of bad fit. If each disjunct is rejected then the composite *SI* ought to be rejected for bad fit. In contraposition, if *SI* ought not to be rejected, then there is at least one disjunct h_t which, taken individually, fits well enough.

Arthur Bowley never accepted Fisher's argument concerning Chi-Squared. In the last (sixth) edition of his text, *Methods of Statistics* (Bowley [1947]), first published in 1900, he argues against Fisher. His argument employs considerations similar to my remarks above about the 'disjunctive criterion for rejecting composite hypotheses'. Bowley is perhaps unique in continuing to resist Fisher's proof. Below I present a reconstructed version of Bowley's argument.

Suppose one computed expected frequencies for each h_i in the disjunction (h_i v . . . v h_ω v . . .) ≡ *SI*. No estimation is necessary since each h_i is simple. For each set of expected frequencies a χ_0^2 value could then be calculated using the observed frequencies. Let $\bar{\chi}$ be the set of χ_0^2 values so determined. While such a procedure could not be carried out in practice, Bowley observes that we do know something about $\bar{\chi}$: the χ_0^2 value obtained when expected frequencies are estimated by fixing the marginals, call it χ_0^{2m}, is the minimum of $\bar{\chi}$. If one then determined probabilities for each $\chi_0^2 \in \bar{\chi}$, using the table for four cells since no estimation is involved, the maximum probability, P^m, would correspond to χ_0^{2m}. Bowley reasons: if P^m is small enough to warrant rejection of the simple hypothesis to which it corresponds, then no disjunct of $SI \equiv (h_i \vee \ldots)$ is close enough to the truth; they all, taken individually, fit badly. Hence, the composite SI fits badly and should be rejected. If, on the other hand, P^m is not small enough to warrant rejection of the simple hypothesis to which it corresponds, then at least one disjunct of SI fits well enough that, taken individually, it would not be rejected. Consequently, since SI is merely a disjunction, it fits well enough not to warrant rejection.

χ_0^{2m} is computed from expected frequencies estimated by fixing the marginal totals. If we adopt the 'disjunctive criterion for rejecting composite hypotheses' then probabilities obtained from the table for four, not two, cells are the right ones to use in testing the composite hypothesis: of independence, SI. Pearson was right!

The Consequence Condition, mentioned at the beginning of the essay, entails the 'disjunctive criterion for rejecting composite hypotheses'. Each disjunct h_i of SI impliesSI since $p \rightarrow (p \vee q)$. By the Consequence Condition, if evidence e warrants the rejection of SI, then e warrants the rejection of h_i for every disjunct h_i. By contraposition, if there is one disjunct h_i such that e does not warrant h_i's rejection, then e does not warrant SI's rejection.

7 GOODNESS OF FIT AND INFORMATION

Chi-Squared applied as Fisher advocated violates the 'disjunctive criterion for rejecting composite hypotheses' and hence the Consequence Condition. Consider the following example: red and black balls are distributed from an urn to two players. We test two hypotheses: 1) h: the manner of distribution gives each possible outcome, {Player A, red; Player A, black; Player B, red; Player B, black}, the same probability; 2) i: ball colour is statistically independent of player identity. h is one simple disjunct of i. The following sample is drawn. The margins are not fixed:

	red balls	black balls	totals
Player A	18	8	26
Player B	8	18	26
totals	26	26	52

To test i, independence, we must estimate exact expected frequencies. If we fix the margins to do so, as Fisher bids, we must use the table for two cells; there is one degree of freedom. The resulting Chi-Squared value, χ_0^2, is 7.69; the probability is 0.00553.

On the other hand, consider a test of h, equal cell frequencies. In this case we compare the evidence directly with the expected cell frequencies of 13 balls per cell. Since there is no estimation there are three degrees of freedom. Everyone agrees in this case: the table for four cells is correct. The resulting Chi-Squared value, χ_0^2, is again 7.69; the probability in this case is 0.05287.

Fisher's methods violate the 'disjunctive criterion for rejecting composite hypotheses'. Applying Fisher's methods we are warranted in rejecting *i* at better than the 0.0056 level of significance; we are not warranted in rejecting *h*, a simple disjunct of *i*, at better than the 0.0056 level of significance. *h* implies *i*, so Fisher's methods also violate the Consequence Condition.

Bowley took this situation to be reason for dissatisfaction with Fisher's treatment of Chi-Squared. I differ. I believe the result shows that Fisher's criteria for good fit differ from solely 'closeness to the truth'; he includes a measure of the informativeness of an hypothesis.

Degrees of freedom measure the informativeness of an hypothesis; the hypothesis, *h*, that a six-sided die is fair has five degrees of freedom; it is a more informative hypothesis than the hypothesis, *i*, that the probability of an ace, two, three or four is 1/6 (the probability of a five or six being unspecified), which has four degrees of freedom. The greater the number of unspecified parameters whose exact values must be estimated from the data, the fewer the degrees of freedom of the hypothesis. An hypothesis with many unspecified parameters—and fewer degrees of freedom—is less informative than one with few or no unspecified parameters.

The stringency of a test of *SI* which accounts for degrees of freedom is less than a test of *SI* which does not account for degrees of freedom. This follows from my reconstruction of Bowley [1947]. Roughly put: the fewer degrees of freedom, the easier it is to reject an hypothesis. Thus, when degrees of freedom are accounted for, the criteria of fit include a measure of the informativeness of an hypothesis, and less informative hypotheses become relatively easier to reject.

Including a measure of informativeness in test construction is appropriate. We seek truth; but we also seek information. A very informative hypothesis not too far from the truth can be better (more adequate) than a less informative hypothesis which is closer to the truth. I argue these points at greater length in my [1982]. Here it is appropriate to point out that Fisher's introduction of degrees of freedom added the criterion of informativeness to measures of fit in particular and tests of statistical hypotheses in general. The result violates the Consequence Condition, but this is not surprising since 'closeness to the truth' is not the only criterion used in test construction.

8 CONCLUSION

I have argued that Fisher's introduction of degrees of freedom to resolve the Chi-Squared Controversy fundamentally altered the sense of 'goodness of fit'. After Fisher an hypothesis is rejected in light of evidence, not because it is likely false, but because it is inadequate. Both closeness to the truth and informativeness are aspects of the adequacy of an hypothesis. It would be incorrect to say that either Pearson or Fisher was wrong; they argue from different points of view regarding statistical tests. It is noteworthy that Fisher's point of view prevailed.

My discussion has focused on the Chi-Squared Controversy and the Consequence Condition. This focus limits the discussion in two important respects: historical detail and analytic generality.

I have left out much of the historical detail. Karl Pearson's position regarding Chi-Squared and contingency tables may not be simply described. He does stress the importance of hypotheses accurately describing observations (Pearson [1960], p. 110 and [1894], p. 2); it is a small step from this emphasis to the 'closeness to the truth' criterion of fit and the 'disjunction criterion for testing composite hypotheses'. However, Pearson's arguments were fundamentally coloured by his high regard for Normal Correlation. A full understanding of Pearson's position requires an understanding of his views on the nature of causality and natural law. Happily for the purposes of this essay, Arthur Bowley's 1947 argument distilled out much of these issues and left exposed the fundamental change Fisher's introduction of degrees of freedom brought about. Here again, Bowley's contributions to the Chi-Squared Controversy were not limited to his 1947 argument. In fact, he communicated Fisher's original 1922 paper to the Royal Statistical Society. A considerably more complete history of the Chi-Squared Controversy may be found in my doctoral dissertation (Baird [1981]).

One might also infer from my discussion that the change in the criteria of fit is restricted to the Chi-Squared test of independence. Such an inference would be completely mistaken; the change is quite general. Fisher's concept of degrees of freedom is applied in the construction of almost any statistical test. As a consequence all of these tests violate the Consequence Condition. A coin flipping example may be found in my [1982]. Indeed one reads in a recent text that few unspecified parameters is one important criterion of a statistical model; another perhaps conflicting criterion is the model's faithfulness to observations (Cox and Hinkley [1974], pp. 5–7). The move Fisher made in the Chi-Squared Controversy has been adopted throughout the theory of statistical inference. The Chi-Squared Controversy does indeed mark a turning point for inductive inference.

University of South Carolina

REFERENCES

BAIRD, D. [1981]: *Significance Tests: Their Logic and Early History*, Ph.D. Dissertation submitted to Stanford University, June 1981.

BAIRD, D. [1982]: 'Information, Falsification and the Consequence Condition', *forthcoming*.

BOWLEY, A. [1947]: *Elements of Statistics*, 6th ed. New York: Charles Scribner's Sons.

COX, D. R. and HINCKLEY, D. V. [1974]: *Theoretical Statistics*. London: Chapman and Hall.

FISHER, R. A. [1922]: 'On the Interpretation of Chi-Square for Contingency Tables and the Calculation of P', *Journal of the Royal Statistical Society*, **85**.

FISHER, R. A. [1959]: *Statistical Methods and Scientific Inference*. New York: Hafner Publishing Co.

HACKING, I. [1976]: *The Logic of Statistical Inference*, 2nd ed. Cambridge: Cambridge University Press.

HEMPEL, C. [1965]: *Aspects of Scientific Explanations and Other Essays in the Philosophy of Science*. New York: The Free Press.

KENDALL, M. and STUART, A. [1977]: *The Advanced Theory of Statistics, Vol. 1: Distribution Theory;* 4th ed. London: Charles Griffin and Co.

MOOD, A., GREYBILL, F. and BOAS, D. [1974]: *Introduction to the Theory of Statistics*; 3rd ed. New York: McGraw-Hill Book Co.

PEARSON, K. [1894]. 'Contributions to the Mathematical Theory of Evolution', page references to the reprint in K. Pearson [1948].

PEARSON, K. [1900]: 'On the Criterion that a Given System of Deviations from the Probable in the Case of a Correlated System of Variables is such that it Can be Reasonably Supposed to have Arisen from Random Sampling', page references to the reprint in K. Pearson [1948].

PEARSON, K. [1904]. 'Mathematical Contributions to the Theory of Evolution: XIII. On the Theory of Contingency and Its Relation to Association and Normal Correlation', page references to the reprint in K. Pearson [1948].

PEARSON, K. [1922]. 'On the χ^2 Test of Goodness of Fit', *Biometrika*, **14**.

PEARSON, K. [1948]: *Karl Pearson's Early Statistical Papers*, edited by E. S. Pearson. London: Cambridge University Press.

PEARSON, K. [1960]: *The Grammar of Science*, reprint of the 3rd (1911) edition. New York: Meridon Books, Inc.

POPPER, K. [1959]: *The Logic of Scientific Discovery*. New York: Basic Books.

YULE, G. U. and GREENWOOD, M. [1915]: 'The Statistics of Anti-Typhoid and Anti-Cholera Innoculations, and the Interpretation of Such Statistics in General', *Royal Society of Medicine Proceedings, Section of Epidemiology and State Medicine*, **8**, part II, pp. 113–194. Also reprinted in Yule [1971]; page references, however, are to the original printing.

YULE, G. U. [1971]: *Statistical Papers of George Udny Yule*, edited by A. Stuart and M. Kendall. London: Charles Griffin and Co.