Statistical Science
1998, Vol. 13, No. 4, 351–367

The 2 × 2 Table: A Discussion from a Bayesian Viewpoint

J. V. Howard

Abstract. The 2 × 2 table is used as a vehicle for discussing different approaches to statistical inference. Several of these approaches (both classical and Bayesian) are compared, and difficulties with them are highlighted. More frequent use of one-sided tests is advocated. Given independent samples from two binomial distributions, and taking independent Jeffreys priors, we note that the posterior probability that the proportion of successes in the first population is larger than in the second can be estimated from the standard (uncorrected) chi-square significance level. An exact formula for this probability is derived. However, we argue that usually it will be more appropriate to use dependent priors, and we suggest a particular "standard prior" for the 2 × 2 table. For small numbers of observations this is more conservative than Fisher's exact test, but it is less conservative for larger sample sizes. Several examples are given.

Key words and phrases: Bayesian statistics, two by two contingency tables, Fisher's exact test, Yates's correction, chi-square tests, significance tests, *p*-values, likelihood principles, conditioning, ancillarity, dependent prior distributions, posterior probability.

1. INTRODUCTION

Suppose that independent random samples are drawn from two large populations, and each member classified as a "success" or a "failure." The first sample is of size n_1 and yields a successes and b failures. The second, of size n_2, yields c successes and d failures. The data is displayed in the familiar 2 × 2 table (see Table 1). It is required to give some idea of the extent to which the data supports (or goes against) the hypothesis that the success proportion in the second population, p_2, is greater than that in the first population, p_1.

This apparently simple problem has given rise to a large literature, stretching back to Karl Pearson's introduction of the chi-square goodness-of-fit test (Pearson, 1900), which he applied to this situation with three degrees of freedom. This was incorrect, because the "expected" values used for the chi-square calculation are computed so as to have the right row and column totals, and thus will fit the "observed" values more closely than if they were simply constrained to the correct overall total. Pearson's error was corrected by Fisher, and the resulting dispute led to a lifelong split between the two men. An interesting survey of the problem (including other variants) is given by Yates (1984). The continuing high level of interest may perhaps be due to the fact that this is one of the simplest natural problems to demonstrate clear differences between classical and Bayesian analyses, and also between different types of classical analysis. It thus forms a sort of "test-bed" for different approaches to statistical inference.

Our general approach will be to imagine that we have been asked to appear before a government committee of nonstatisticians. They wish to be informed whether the data favours H_1: $p_2 < p_1$ or H_2: $p_1 < p_2$, and to be given a quantitative measure of the strength of the evidence in support of the more likely hypothesis. Everyone is certain that p_1 and p_2 will not be exactly equal, and that neither will be 0 or 1. For example, they might be interested in whether English or Scots cattle herds have a higher proportion of cows infected with a certain virus. They are sure that in each population some (but not all) cows will have the virus; and from experience with other diseases they are sure that the

J. V. Howard is Senior Lecturer, London School of Economics, Houghton Street, London WC2A 2AE (*e-mail: j.v.howard@lse.ac.uk*).

TABLE 1
Data in the 2×2 *table*

	Successes	Failures	Total
Sample 1	a	b	n_1
Sample 2	c	d	n_2
Totals	m_1	m_2	N

two proportions will not be exactly equal. (So we are given that p_1 and p_2 are neither 0 nor 1, and that $p_1 \neq p_2$.) Note that there is no natural null hypothesis.

There are of course a number of variations on this problem. Instead of two populations, we might have two treatment groups (a comparative trial). Suppose, for example, we wish to know which of two treatments has a higher success rate. We are sure that without any treatment the success rate is zero, and that neither treatment has a 100% success rate, although both sometimes succeed. Since the treatments have different bases, we are sure there will be a difference in the success rates, but we are not sure in which direction. Neither treatment is established as standard, so again there is no natural null hypothesis. In this variant, instead of random samples from two populations, we take a single random sample from a population of "experimental units," and then make a random allocation of treatments among the units.

If one of the treatments was a placebo and the other was (say) a homeopathic remedy, it might be reasonable to test the point null hypothesis of no treatment effect at all. Similarly, if we were testing for extrasensory perception we might regard the hypothesis of zero effect as being very possibly exactly true. In this paper, we do not consider situations of this type where a precise null hypothesis is to be tested (but see Berger and Delampady, 1987, and Berger, Boukai and Wang, 1997).

Two other alternative situations will not be considered in detail. First, we could have a single random sample of size N from one population, each member of the population being classified in two different ways (a double dichotomy). The parameters for this problem are the probabilities p_{11}, p_{12}, p_{21}, p_{22} of the four cells. Second, there is the "lady tasting tea" discussed by Fisher in *The Design of Experiments* (Fisher, 1935). Here, we again have a random sample of experimental units (in this case, cups of tea), and as before we make a random allocation of treatments (preparing the tea by pouring the milk first or last). However, the taster is told how many cups were given each treatment, so when she classifies them she will definitely get the correct total numbers in each group. In this case the model might involve two parameters: p_1 and p_2, where p_i is the probability of guessing correctly when presented with a single cup which was given treatment i (or it might have just one parameter p if we assumed $p_1 = p_2$). These variations differ from our problem in the amount we know about the marginal totals in the table (m_1, m_2, n_1 and n_2) before the data is collected or the experiment performed. Table 2 summarizes the four possibilities.

In order to prepare our evidence to the committee, we begin by reviewing briefly in Section 2 the general ideas available from classical and Bayesian statistics. Then in Section 3 we take a particular table and ask what sort of statement different types of statisticians might make to the committee. Our general approach will be that, after the data has been collected, a statistician can make various possible *hypothetical* statements to indicate the strength of the evidence in favor of one hypothesis as against the other. Each hypothetical statement is intended to be unarguably correct (given the model and all the background assumptions). So a statistician of any persuasion would accept all of the hypotheticals as true, but he might feel that some of them were completely irrelevant to the question being addressed. A classical statement, for example, might begin "if H_1 were true, and if the experiment were repeated many times . . . ." A Bayesian might start "if the prior were $f(p_1, p_2)$, then the posterior would

TABLE 2
Classification of 2×2 *table variants*

Type	Number of populations	Parameters	Procedure	Number of margins fixed
Double dichotomy	1	$p_{11}, p_{12}, p_{21}, p_{22}$	Take random sample from the population; classify in two ways	0
Two binomials	2	p_1 and p_2	Take a random sample from each population; classify	1
Comparative trial	1	p_1 and p_2	Take a random sample from the population; make a random allocation of treatments	1
Tea tasting	1	p_1 and p_2 or $p = p_1 = p_2$	Take a random sample from the population; make a random allocation of treatments	2

be . . . ." Of course the Bayesian should first make the indicative statement "my prior was . . . , so my posterior is . . . ," but when she wanted to communicate her results to other statisticians (even other Bayesian statisticians) she would be likely to use some form of standard prior, or else give results for a whole range of priors.

Sections 4 and 5 discuss some problems with the classical approaches (including stopping rules and two-sided tests). Then we attempt some sort of reconciliation of the different calculations. It is obviously convenient if the same computation can have both a classical and a Bayesian interpretation (i.e., can be used to give the numbers to be fed into both classical and Bayesian hypothetical statements). We observe in Section 6 that the usual uncorrected chi-square test can be regarded as giving an approximation to a Bayesian posterior probability that $p_2 < p_1$. However, we have to assume *independent* prior beliefs about the two populations in order to get this result, and we feel this will often not be a reasonable approximation to the statistician's prior. Consequently we introduce in Section 7 a conjugate family of priors which incorporate *dependence* between beliefs about the two populations. It appears that fairly mild beliefs about dependence can make our posterior statements much more cautious for small sample sizes. We compare various amounts of prior dependence and suggest one particular prior as a "standard" dependent prior. Although a Bayesian should use her own prior, she may also need to exhibit results from a range of priors for the benefit of other statisticians, and the inclusion of one (or more) standard priors in this set makes it easier for others to judge the strength of evidence in a given body of data.

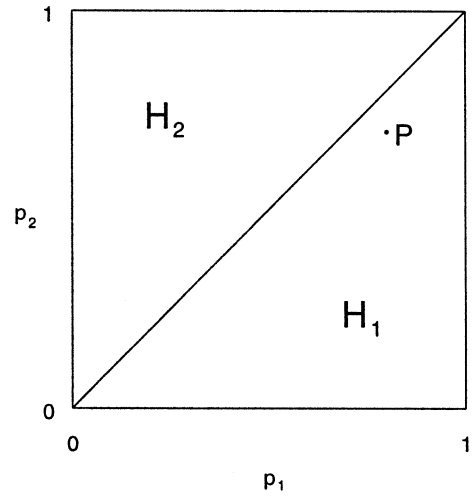

FIG. 1. *The unknown population parameters are shown by the point P with coordinates* (p_1, p_2) *in the unit square (parameter space):* p_1 *is larger than* p_2 *in triangle* H_1*, and smaller than* p_2 *in* H_2.

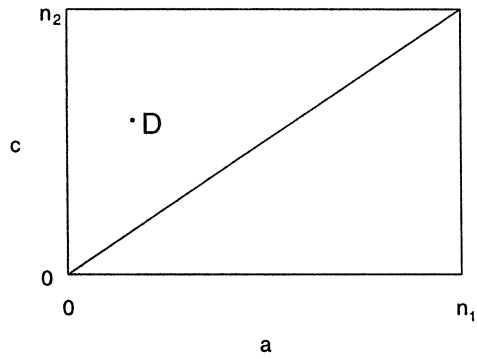

FIG. 2. *The sample space consists of all pairs* (a, c)*, where a is the number of successes in population* 1 *and c the number in population* 2*. The observed data can be shown as the point D in this space.*

2. INITIAL REVIEW

In the two binomials problem the unknown state of the world can be represented by a point $P = (p_1, p_2)$ in the unit square (see Figure 1).

The data can be shown (Figure 2) as the point $D = (a, c)$ in the rectangle with dimensions (n_1, n_2). This rectangle could be rescaled to the unit square and D would then become the point $\hat{P} = (\hat{p}_1, \hat{p}_2)$, where $\hat{p}_1 = a/n_1$ and $\hat{p}_2 = c/n_2$ are the observed success proportions in the two populations; $\hat{P}$ is an estimate of the unknown P.

Now our problem is this: given D (a specific point in the upper triangle of Figure 2), how strong is the evidence against P's lying in the lower H_1 triangle of Figure 1? There are three main non-Bayesian approaches to this problem.

2.1 A Likelihood Approach

A good starting point is to look at the likelihood function, namely, the probability of getting the data D as a function of the positions of P. This is

$$l(p_1, p_2) = p_1^a(1 - p_1)^b p_2^c(1 - p_2)^d.$$

We could, for example, look at the ratio of the maximum likelihood we can achieve when P is unconstrained to the maximum that can be achieved when P is constrained to the H_1 triangle. Contours of one particular scaled likelihood function (for the numerical example discussed in Section 3) are shown in Figure 3. In this example $a = 3$, $b = 15$, $c = 7$ and $d = 5$, and the figure plots contours of $1.5 \times 10^8 l(p_1, p_2)$. (The diagram shows that a likelihood ratio greater than 9 can be achieved in this case.)

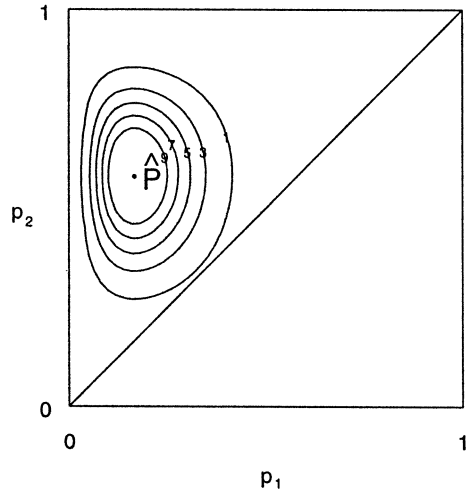

FIG. 3. *The likelihood function $l(p_1, p_2)$ achieves its maximum at $\hat{P}$ with coordinates $(a/n_1, c/n_2)$. Contours of $1.5\times10^8 l$ are plotted. There is a likelihood ratio of at least 9 between points within the smallest contour around $\hat{P}$ and points outside the largest contour.*

This approach holds resolutely to the idea that we should take into account the probability of what was observed, but definitely not the probabilities of things that were not observed. Edwards (1972) based his approach to statistical inference on likelihood, but used on its own it can lead to problems (see Fraser, Monette and Ng, 1984, and the discussion in Goldstein and Howard, 1991). However, the Bayesian approach would use the likelihood function to weight the prior distribution over the unit square in Figure 1, and then renormalize to obtain a posterior distribution. In fact, with a uniform prior Figure 3 would show contours of the scaled posterior—the posterior would be proportional to $l(p_1, p_2)$. Integrating the posterior over the lower triangle gives the Bayesian posterior probability that H_1 is true, which is the sort of number we are seeking. With a uniform prior, this probability is

$$\frac{\int_{p_1=0}^{1}\int_{p_2=0}^{p_1} l(p_1, p_2)dp_2\,dp_1}{\int_{p_1=0}^{1}\int_{p_2=0}^{1} l(p_1, p_2)dp_2\,dp_1}.$$

Provided the prior is proper, the Bayesian approach does not seem to lead to any paradoxes.

2.2 Frequentist Inference

Classical statisticians go in the opposite direction. Instead of keeping D fixed and allowing P to vary, they would tend to look first at a fixed P in the H_1 triangle of Figure 1, and then calculate the probability of observing a point in some "extreme region" of the upper triangle of Figure 2 which just includes D. This is a p-value calculation. (The idea is basically that either H_1 is false or an event with a surprisingly low probability has occurred.) For example, we could take as a measure of the evidence in favor of H_2 $(p_2 > p_1)$ over H_1 $(p_2 < p_1)$ the difference between the proportion of successes in the two populations $(\hat{p}_2 - \hat{p}_1)$. Let this "test statistic" be the random variable V:

$$V = \hat{p}_2 - \hat{p}_1 = \frac{c}{n_2} - \frac{a}{n_1}.$$

Suppose the value of V at the data point D is $V(D) = v$. We can now calculate the probability that if the experiment were repeated (for some fixed P) we would obtain a new value for $\hat{p}_2 - \hat{p}_1$ as large or larger than v. Figure 4 shows the line $V = v$ through D (parallel to the main diagonal of the rectangle) and shades the "critical region" where $V \geq v$.

In fact $\hat{p}_2 - \hat{p}_1$ is never used as a test statistic in this simple form. Instead it is standardized in some way. Very commonly it is rescaled by dividing by the standard deviation it would have if $p_1 = p_2 = m_1/N = (a + c)/(a + b + c + d)$, i.e., the total number of successes in the two populations divided by the total number of trials, the pooled estimate of a common success proportion. The point $P = (m_1/N, m_1/N)$ lies on the diagonal of parameter space (Figure 1), but when it is rescaled it can be plotted in sample space as the point E in Figure 5, which is the point on the diagonal of the rectangle such that DE has slope -1. This estimation of a common success proportion by pooling leads to Yule's test statistic

$$Z = \{ad - bc\}\sqrt{\frac{a + b + c + d}{(a + b)(c + d)(a + c)(b + d)}},$$

which gives a critical region similar to that shown in Figure 5. (The diagrams continue to be based

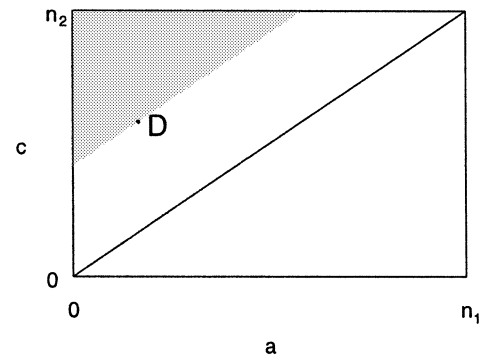

FIG. 4. *One possible critical region (shaded) would consist of all combinations of a and c for which $V = \hat{p}_2 - \hat{p}_1 = c/n_2 - a/n_1$ is greater than or equal to $V(D)$, the value for $\hat{p}_2 - \hat{p}_1$ found in the experiment.*

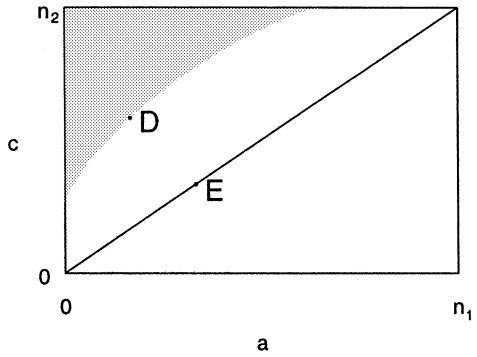

FIG. 5. *The shaded region shows the points in the upper triangle of the sample space for which Pearson's chi-square statistic is greater than or equal to the observed chi-square value at D; E is the point in sample space which has the same number of successes in total $(a+c)$ as D but having equal proportions of successes in the two populations.*

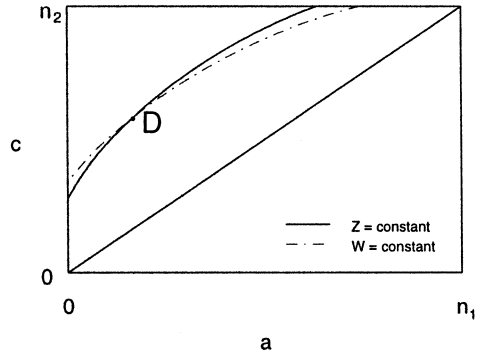

FIG. 6. *Boundaries of the critical regions which just include D for the two test statistics Z and W.*

on the numerical example discussed in Section 3.) Yule's statistic is in fact simply the signed square root of Pearson's (uncorrected) chi-square statistic for the 2×2 table.

Another possibility is to calculate a confidence interval for $p_2 - p_1$ and see whether the interval includes zero. This means that we standardize $\hat{p}_2 - \hat{p}_1$ by dividing by the standard deviation it would have if in fact $p_1 = \hat{p}_1 = a/n_1$ and $p_2 = \hat{p}_2 = c/n_2$ (i.e., if P was in fact $\hat{P}$). So we do not estimate a common success proportion by pooling, but instead make a separate estimate for each population. This leads to the unpooled test statistic

$$W = \{ad - bc\}\sqrt{\frac{(a+b)(c+d)}{ab(c+d)^3 + cd(a+b)^3}}.$$

Since we will be making exact calculations, all that matters is the ordering of the points in the sample space imposed by the test statistic. Suissa and Shuster (1985) noted that when the two sample sizes are equal $(n_1 = n_2)$ Z and W are increasing functions of each other, and so give the same ordering of the sample space. Hence in this case pooling and not pooling are equivalent when an exact calculation is made. (Robbins, 1977, had asked which was more powerful.) However, when the sample sizes are unequal the orderings may differ: Figure 6 shows the two boundary lines for the critical region in our example (the dashed line is the unpooled boundary). When the two orderings are different, it is easy to see that for some data points one ordering will give a smaller p-value, while for other data points the other ordering will give the smaller number. Barnard (1947) discussed desirable attributes for a logical ordering. In this paper we will use the ordering induced by Yule's statistic.

Even after deciding on the exact boundary of the critical region, there is still the problem of which point P in the H_1 triangle to use to calculate the probability that the extreme event will occur. We will usually find the location for P which maximizes the probability of the critical event. This will never be found in the interior of the H_1 triangle, but will always occur on the boundary of the triangle, the diagonal line $p_2 = p_1$. Figure 7 shows some contour lines along which P gives equal probability to the critical region of Figure 5—the region where $Z \geq Z(D)$. So the location for P that we are seeking will be found along the contour line of highest probability which just touches $p_2 = p_1$ (this gives the highest probability that can be achieved when P is restricted to the H_1 triangle, because contours

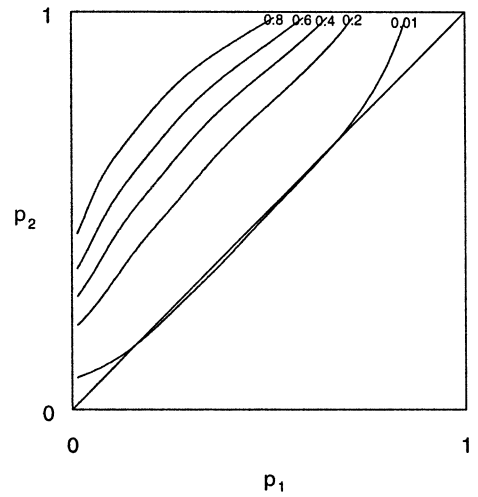

FIG. 7. *The probability that $Z \geq Z(D)$ is a function ζ of p_1 and p_2: ζ is 1 when $(p_1, p_2) = (0, 1)$ and 0 when $p_1 = 1$ or $p_2 = 0$. Contour lines of ζ are plotted for $\zeta = 0.01,\ 0.2,\ 0.4,\ 0.6$ and 0.8. It is just possible for ζ to be larger than 0.01 for values of $p_1,\ p_2$ in the H_1 triangle.*

of higher probability will lie entirely within the H_2 triangle).

2.3 Conditional Inference

However, many classical statisticians follow Fisher (1945) in thinking that the appropriate probability should be conditional on the observed marginal totals m_1 and m_2. This means that we condition on D lying on the straight dashed line with slope -1 shown in Figure 8. Two arguments are given to support this conditional inference approach.

First, we are invited to compare this problem to one where we are interested in a single binomial parameter p, and we make m independent trials with probability p of success. However, m is first sampled from a distribution with unknown parameter q. We know of no relationship between p and q. So our data is the number of successes s and the sample size m, and the parameters are p (of interest) and q (nuisance). The likelihood can be written as

$$L(s, m \mid p, q) = \phi(s \mid m, p)\psi(m \mid q).$$

Note that ψ has no dependence on p and q is not involved in ϕ. In this situation—where the sample size m is (partial) ancillary for p (does not depend on p)—it is argued that we should make the same inferences about p as when m is fixed and given from the start (see Lehmann, 1986, and Cox and Hinkley, 1974); m affects the informativeness of the experiment, but gives no information about p.

It is next suggested that we look at the parameters of our problem as (say) the odds ratio

$$r = \frac{p_1(1 - p_2)}{(1 - p_1)p_2}$$

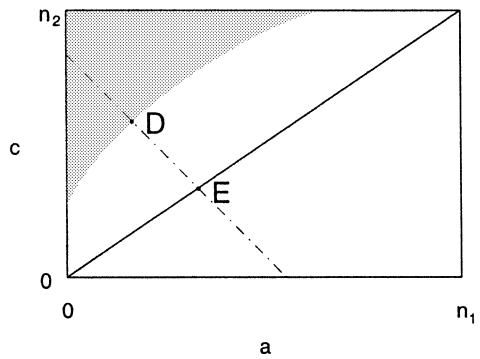

FIG. 8. *The line with slope -1 through D corresponds to data having the same total number of successes (i.e., corresponds to a set of 2×2 tables with the same marginal totals as for the observed data D).*

(of interest) and the odds product

$$t = \frac{p_1 p_2}{(1 - p_1)(1 - p_2)}$$

(nuisance), with data m_1 (showing the informativeness of the experiment—values of m_1 near 0 or N being uninformative) and a (directly relevant); r is of interest because it is a measure of the discrepancy between p_1 and p_2, whereas t measures a sort of average of p_1 and p_2 ($\sqrt{t}$ is the geometric mean of the odds of p_1 and p_2). We would now hope to be able to write the likelihood in the form

$$L(a, m_1 \mid r, t) = \phi(a \mid m_1, r)\psi(m_1 \mid t).$$

Unfortunately, however, no matter how we reparameterize, m_1 is not an ancillary statistic; ϕ—the conditional distribution of a given the column totals m_1 and m_2—is indeed a function of r alone, with no dependence on t. Although ψ—the distribution of m_1—depends mainly on t, it also has some dependence on r. The argument therefore has to be made approximate: m_1 contains very little information about r, so we should treat it *as if* it were ancillary.

Another consequence of this is that a Bayesian approach which put independent priors on r and t and then calculated a posterior for r using just the conditional likelihood ϕ would not be exactly correct. (Cornfield, 1956, used the fact that ϕ depends only on r to calculate exact and approximate confidence limits for r corresponding to the use of Fisher's conditional test.)

The second argument for conditioning looks at Figure 7 and notes that some points in the H_1 triangle give a higher probability to the critical region than other points in the H_2 triangle. (See the 1% probability contour.) Assuming that this is undesirable, is it possible to alter the critical region so as to make the diagonal a contour line? Tocher (1950) showed that this can be done, provided that the test does not depend only on the observed data (a, c), but also on the result of a separate independent random draw from a uniform distribution on $[0, 1]$. (This gives the UMP unbiased test of $p_2 = p_1$ against $p_2 > p_1$.) This randomized test for a given significance level, α, is in fact a conditional test with the randomization used to achieve the exact α required — so we use the random number only when the data point D is the first point outside the unrandomized level-α conditional critical region as we move along the line through D with constant m_1. This result strengthens the argument for a conditional test, even though many would baulk at making a separate randomization before declaring significance.

TABLE 3
Pearson's example

	Successes	Failures	Total
Sample 1	3	15	18
Sample 2	7	5	12
Totals	10	20	30

We will draw on these classical and Bayesian ideas in order to construct various possible hypothetical statements that we might make to our committee members.

3. HYPOTHETICAL STATEMENTS

Supposing that the data has been collected, we can now consider various possible statements which could be made to indicate the strength of the evidence in favor of one hypothesis and against the other. If the general model of the situation is accepted, there is no argument about the accuracy of these statements, only about which statement is most appropriate. Let us illustrate with a particular example from Egon Pearson (1947). The data, given in Table 3, would certainly seem to suggest fairly strongly that there is a higher proportion of successes in population 2 than in population 1. But how strongly? Here are some statements that could be made to the committee after collecting the data.

THE UNCONDITIONALIST. If in fact H_1 is true ($p_2 < p_1$) and if the experiment were repeated independently, once again sampling 18 subjects randomly from population 1 and 12 randomly from population 2, the probability that we would achieve a result as extreme as or more extreme than this (and in the same direction) would be at most 1.23%.

1. An alternative phrasing often used in textbooks speaks of "repeating the experiment many times" giving a proportion of times with an equally or more extreme result of 1.23%. The second statement simply results from applying the strong law of large numbers and using the first statement.
2. We have taken the "extremity" of a result as being determined by the value of its uncorrected chi-square statistic as discussed in Section 2, giving the critical region shown in Figure 5. This figure is somewhat misleading because the problem is discrete. Figure 9 shows the finite set of extreme outcomes (marked with asterisks). We can notice that although the data point D (namely, the point $(3, 7)$ with $m_1 = 10$) has a Yule statistic of 2.37, the other boundary points have higher values. For example, if $m_1 = 9$, the least extreme point is $(2, 7)$ with Yule statistic 2.77, and if $m_1 = 8$, it is $(1, 7)$ with Yule statistic 3.20.
3. In practice the unconditionalist usually does not bother to calculate the exact probability of extreme results. Normally he takes the uncorrected chi-square tail probability as a reasonable estimate: in this case he would quote 0.89%.
4. The unconditionalist imagines repeating the experiment with the same sample sizes, so he does condition on the row totals, although not on the column totals.

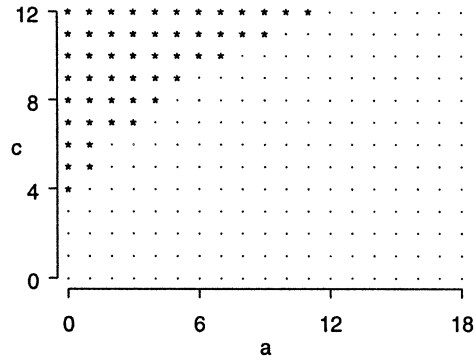

FIG. 9. *Points in the sample space for Pearson's example: points in the critical region of Figure 5 are shown as asterisks.*

The statement made by the unconditionalist should be compared to the following statement.

THE CONDITIONALIST. If in fact H_1 is true ($p_2 < p_1$) and if the experiment were repeated independently, once again sampling 18 subjects randomly from population 1 and 12 randomly from population 2, the probability that we would achieve a result as extreme as or more extreme than this (and in the same direction) conditional on having the same column totals (10 and 20) would be at most 2.42%.

1. Again, we could think of repeating the experiment many times and discarding all the results except those having the correct column totals. Then if H_1 is true, the proportion of the retained results which is as extreme as or more extreme than the data would be at most 2.42%.
2. This time there is no doubt about the extremity of a result: all the results considered lie on the line shown in Figure 10. Moreover, the conditional probability distribution over these points is the same whenever $p_2 = p_1$.
3. Some statisticians prefer to include only half of the probability of the data point D actually

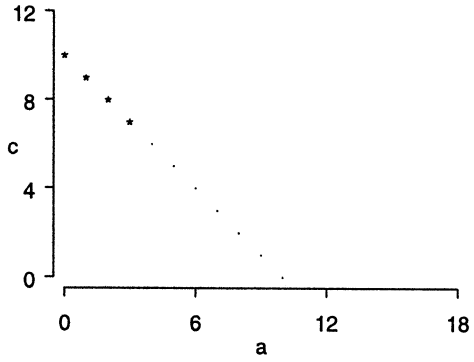

FIG. 10. *Critical points on the conditional line.*

observed. This has the effect of making the expected p-value exactly 0.5 when $p_2 = p_1$ (see Lancaster, 1961; Stone, 1969; Franck, 1986; Haber, 1986; and Routledge, 1992). This would make our hypothetical statement more complicated, and reduce the quoted probability from 2.42% to 1.34%.

4. In practice the conditionalist might not calculate the exact probability using the formula for Fisher's "exact" test, but instead use Yates's correction to the standard chi-square statistic. In this case he would then quote 2.41%.
5. The conditionalist calculates a larger probability than the unconditionalist because the least extreme points on the other diagonals have larger test statistics than D (see note 2 for the unconditionalist statement).

Suppose the data came from a comparative trial. Some classical statisticians prefer in this situation to make a statement based on the random allocation of treatments over subjects. This is the Fisher randomization test:

THE FISHERIAN. Suppose there was in fact no difference at all between the treatments (not just that their success rates were equal, but so that any individual would be cured by treatment 1 if and only if he would be cured by treatment 2). Then if the experiment were repeated with the same 30 subjects (assuming this were possible), randomly allocating 18 to treatment 1 and 12 to treatment 2, the probability of obtaining a result as extreme as or more extreme than the one obtained (and in the same direction) would be exactly 2.42%.

1. The calculation here is the same as for the conditionalist, but the interpretation is different. This type of statement cannot be made unless we have different experimental treatments. It is also difficult to see how this statement helps to make any inference about the populations in general, as it refers only to 30 particular individuals (who, as far as this statement is concerned, might not have been randomly sampled from the population).
2. Even as hypotheticals go, this is far more hypothetical than most. We could actually sample a new set of 30 subjects and assign them treatments at random, but in most cases it would not be possible to reallocate the treatments already given to the original 30 subjects.

A Bayesian hypothetical will take the form "if the prior were this, the posterior would be . . . ." For definiteness, let us take a Jeffreys prior

$$f(p_1, p_2) \propto p_1^{-1/2}(1 - p_1)^{-1/2} p_2^{-1/2}(1 - p_2)^{-1/2}.$$

The posterior will then be proportional to

$$p_1^{a-1/2}(1 - p_1)^{b-1/2} p_2^{c-1/2}(1 - p_2)^{d-1/2}.$$

This can then be integrated over the H_1 triangle, either exactly (see the Appendix), or numerically. We then have a Bayesian statement:

THE BAYESIAN. Suppose p_1 and p_2 were sampled from independent Jeffreys priors. After the data was observed the posterior probability that $p_2 < p_1$ would be 0.87%.

1. Bayesians might reasonably complain that their statements are not hypothetical: the statistician should say "my prior *was* . . . so my posterior *is* . . . ." However, when the results are to be communicated to others, whose priors may be unknown to the author, the statements must become hypothetical. Of course, it would be good practice to exhibit several calculations to illustrate the effect of the data on different prior beliefs.
2. We show in the Appendix how to calculate the posterior probability (0.8723%) exactly when starting with a Jeffreys prior. The fact that this number is very close to the p-value obtained from the standard Pearson chi-square statistic (0.8853%—see the unconditionalist, note 3) is not a coincidence. We show in Section 6 that the uncorrected chi-square statistic gives an approximation to the Bayesian posterior for exactly the same reasons that the corrected (Yates) chi-square statistic gives a good approximation to the Fisher exact probability (see the conditionalist, note 4).
3. For comparison, if we used a Laplace (independent uniform) prior, the posterior probability

would be 1.07%. With a Haldane (independent improper) prior

$$f(p_1, p_2) \propto p_1^{-1}(1-p_1)^{-1}p_2^{-1}(1-p_2)^{-1},$$

the posterior probability would be 0.69%. The fact that these numbers are all much smaller than the Fisher test value of 2.42% might lead us to think that the exact test is far too conservative (as has often been argued). We shall later dispute this and claim that with more reasonable priors the Fisher value is by no means too large.

4. Although the "half" prior is usually referred to using Jeffreys's name, he himself preferred the uniform prior in the absence of information (see Jeffreys, 1961, page 125).

4. DISCUSSION OF THE CLASSICAL STATEMENTS

There is a strong analogy between the problems arising in classical statistics and those occurring in frequentist probability theory. Consider the unconditionalist. He envisages an infinite sequence of experiments in which a definite proportion (1.23%) give a particular result (namely, a result as extreme as or more extreme than the one originally observed). Compare this to the classical explanation of an event having probability 1.23%—the event must be regarded as a member of an infinite sequence of trials in which the successful trials occur with limiting frequency 1.23%. Such a sequence was called a *collective* by von Mises, who required that the same limiting frequency should occur in any subsequence of the original sequence chosen by an arbitrary rule. Unfortunately, there will always be some rule which selects precisely the successful trials, so this cannot work. Church (1940) made the definition precise by taking "rule" to mean "computable algorithm" and showed that with this definition collectives did indeed exist.

Suppose we do repeat the experiment many times, as envisaged by the unconditionalist. Then we have a case where the event (namely, a result as extreme as or more extreme than the one originally observed) belongs to a sequence with limiting frequency 1.23% *and* to a subsequence (experiments giving the same column totals) with limiting frequency 2.42%. Which number should we quote? There is a strong argument that when the conditioning event (here having the given column totals of 10 and 20) gives no information about the parameters of interest, we *should* condition. For example, Cox and Hinkley (1974) state "there are serious difficulties in a sampling theory approach that does not take account of a conditionality principle, ..."; Lehmann (1986) says "... if repetitions ... are potential rather than actual, interest will focus on the particular event at hand, and conditioning seems more appropriate." However, in the case of the 2×2 table the column totals do contain some information about $p_2 - p_1$. To see this, consider a Bayesian whose prior was symmetric about the line $p_2 = p_1$. If m_1 was ancillary, then informing her of the value of m_1 would not alter her probability that H_1 was true from its initial value of 1/2. However, if we calculate the posterior probability that $p_1 < p_2$ for the Pearson example with independent uniform priors, supposing that we have been informed of the row and column totals (18, 12 and 10, 20) but not knowing the values in the body of the table, we find a conditional probability of 59%. This does not seem to be a negligible increase from the original 50%. (Although a preliminary observation of one failure in population 1 or one success in population 2 would have shifted the prior probability more—to 67%.) The problem is: when the conditioning event contains some (but not a great deal of) information about the parameters of interest, should we use the conditional or unconditional probability in our hypothetical statements? Barnard, who devised the unconditional CSM test (Barnard, 1947) was converted by Fisher to the conditional test, and in turn has recently converted Upton (see Upton, 1992). Nevertheless, the argument is by no means settled. The discussion following Yates's 1984 paper seemed mostly in favor of conditioning, but since then Suissa and Shuster (1985) and D'Agostino, Chase and Belanger (1988) have put the unconditional argument. Earlier advocacy comes from Grizzle (1967), Conover (1974), and Berkson (1978). Little (1989) puts the opposing view. (For testing a precise hypothesis, Berger, Boukai and Wang, 1997, try to relate conditional frequentist inference to Bayesian methods.)

Further difficulties arise when an event can be regarded as a member of several possible sequences or subsequences. Imagine that coins are produced by a mint and tossed repeatedly. The coins wear over time and so does the mint's mechanism. Consider the 10th toss of the 10th coin produced by the mint. We could regard this as a member of the sequence "tosses of the 10th coin," or of the sequence "10th tosses of any coin," or indeed of the sequence "n'th tosses of the n'th coin." All these sequences may have different limiting frequencies. We might avoid this problem in the foundations of probability theory by working with the idea of a *model* which predicts a probability p_{ij} for the ith toss of the jth coin. Whether the model is acceptable has then to

be decided by statistical theory. So the problem is relocated from the foundations of probability theory to the foundations of statistical theory.

Unfortunately, however, the same sort of difficulty reappears in statistical theory in the guise of the notorious stopping rule paradox. In our example, suppose the data had arisen by sampling population 1 until three successes were observed and population 2 until seven were seen. The same 2×2 table is now regarded as a member of a different infinite sequence, and in this sequence it may occur with a different limiting frequency. In the example, if in fact $p_1 = p_2 = 1/3$, the probability of getting our table or a more extreme one (by the chi-square criterion) with this sampling scheme is 0.90%, whereas for the original sampling scheme it would be 1.23%. We could even imagine two experimenters collaborating on the experiment. One believes that they are in the fixed sample size (binomial) situation, the other that they are sampling to get specified numbers of successes. They both agree to stop the experiment at the same point, but then discover that they are calculating different p-values!

If we do have this sampling scheme (two negative binomials instead of two binomials) we do not just get a new unconditional statement. We also get new conditional and Fisherian calculations. (The Bayesian hypothetical is unchanged because the likelihood is the same.) The conditionalist could argue that the total number of failures, m_2, is uninformative and condition on that. This leads to calculating a hypergeometric tail probability, equivalent to applying the exact test to the table $(a - 1, b; c, d)$. For our example this gives a p-value of 1.16%.

The Fisherian will arrive at different answers depending on the randomization method used to allocate treatments, and on the particular sequence of successes and failures observed. Suppose, for example, that we allocate treatments by drawing balls from an urn. Initially the urn has 3 blue balls and 7 green. When a patient arrives we draw a ball and allocate treatment 1 if the ball is blue, 2 if it is green. If the treatment succeeds we do not replace the ball, but if it fails we do. So we will eventually get 3 successes for treatment 1 and 7 for treatment 2. Now suppose further that we observe 20 failures (15 with blue and 5 with green) and then 10 successes (3 with blue and 7 with green). Assuming the treatments had no effect on the sequence of successes and failures, the Fisherian would calculate the probability of drawing a sequence of balls which would give a result as or more extreme in favor of H_2. This is simply the probability of getting 15 or more as a sample from Bin(20, 0.3). The tail probability is vanishingly small—0.004%. For a different sequence, consider 8 successes, then 20 failures, then 2 successes. We need to calculate the probability of getting either 7 greens and 1 blue in the first 8 draws or 6 greens and 2 blues followed by 15 or more blues from the next 20 draws. This gives a Fisherian p-value of 3.19%.

There are in fact a large number of possible sampling schemes. We can sample with fixed n_1 and n_2 (which we imagine is what actually occurred), or until we get given values of a and c, or with a and n_2 given (a binomial–negative binomial situation), or in the case of a double dichotomy until we get a given value in one particular cell (the total sample size then being random); and each sampling procedure can generate an unconditional, conditional and Fisherian statement. The possible ramifications seem extensive.

One experiment that was actually carried out with a different stopping rule is described in Bartlett et al. (1985) and discussed in Cornell, Landenberger and Bartlett (1986), Wei (1988) and Begg (1990). Patients were allocated one of two treatments effectively according to the color of a ball drawn (with replacement) from an urn. The urn started with one ball of each color. If the selected treatment was a success, another ball of that color was added to the urn before the next draw. If the treatment was a failure, a ball of the other color was added. The experiment ended as soon as 10 balls had been added of the same color. This is a randomized play-the-winner rule (see Wei and Durham, 1978). In the event, the outcome sequence was a success for treatment A, a failure for B, then eight more successes for A. If we argued that we have simply observed nine trials of A, all successes, and one of B, a failure, the exact test on the table $(9, 0; 0, 1)$ would give a one-sided p-value of 10%; but other tests are possible, taking account of the stopping rule, and giving different answers. Alternatively, we could use the Fisherian argument: suppose that there was no difference between A and B, and that whatever treatments we applied we would have seen the same sequence of successes and failures $(S, F, S, \ldots, S)$ with any sequence of colors. The observed sequence of treatments $(A, B, A, \ldots, A)$ is in fact the most extreme possible in favor of A, given the success–failure sequence. Hence we calculate a one-sided p-value of

$$\frac{1}{2} \times \frac{1}{3} \times \frac{3}{4} \times \frac{4}{5} \times \cdots \times \frac{10}{11} = \frac{1}{22} = 4.55\%.$$

So this is another example where different arguments give quite different p-values.

For yet another example, suppose we had sampled 30 patients and given each a 60% chance of receiving treatment 1 and a 40% chance of treatment 2. Our data could have arisen from this experimental protocol (a rather poor one as it happens). The likelihood principle says that we should draw the same inference from the data in all these situations. We would agree. All these designs can be looked at as trees, where a path through the tree is determined by the outcome of some random event at each node. When the experiment is repeated we can branch from the original path to a completely different part of the tree, with completely different chance nodes. We would instead suggest thinking of repetitions in which we simply retry each of the nodes visited in the original experiment, ignoring the fact that with the original protocol our results might cause us to branch from the original path. So, just as the actual experiment determines the critical region for the imagined repetitions, we would also allow it to determine the sequence of trials made in those repetitions, even when the imaginary results of these trials would have caused a branching to different nodes of the tree. This argument effectively means that we would regard all these designs as equivalent to one with two fixed sample sizes, the sizes realized in the original experiment. (It is closely related to Dawid's prequential principle; see Dawid, 1984.) So we can see no clear argument for introducing new hypotheticals even when the data comes from one of these more exotic designs.

5. ONE-SIDED OR TWO-SIDED TESTS

Another disputed matter is whether to make one-sided or two-sided statements. We deliberately described the situation as one in which neither H_1 nor H_2 could be singled out as the null hypothesis. Nevertheless we have calculated "significance levels" as if H_1 (the hypothesis not favored by the data) were the null hypothesis, and as if we were making a one-sided test against the alternative H_2. It could be argued that as we have no prior reason to prefer either hypothesis we should always make two-sided statements, talking about the "probability of obtaining a more extreme result in either direction when $p_1 = p_2$." Why have we not done this?

The reason is mainly that we are excluding the possibility that p_1 is exactly equal to p_2. We are definitely assuming imprecise hypotheses (contrast this with Berger and Delampady, 1987). Using an idea advanced by Pratt (1965), we make the same sort of argument for our one-sided statements as that usually made for confidence intervals, in the following way. Suppose the experiment were repeated many times (let us say, to be precise, in the way envisaged by the unconditionalist) and after each experiment we stated the following:

either (i) "H_1 is true: $p_2 < p_1$";
or (ii) "H_2 is true: $p_1 < p_2$";
or (iii) "there is not enough data to decide".

We make statement (i) when the data favors H_1 and when the chi-square value is at least 5.625 (the value for our example). Similarly we make statement (ii) when the data favors H_2 and when the chi-square value is at least 5.625. If the chi-square value is less than 5.625, we make statement (iii). Now, what is our maximum error rate?

If in fact H_1 is true, we make a false statement only when we state (ii), which will occur with a maximum probability of 1.23% when $p_1 = 35.3\%$ and p_2 is slightly less than this. Figure 7 shows the basis of the calculation: along the diagonal $p_2 = p_1$, the highest probability reached is 0.0123 when $p_1 = p_2 = 0.353$. (This is the calculation made by the unconditionalist.) If in fact H_2 is true, the maximum error rate must be exactly the same (occurring when $p_1 = 64.7\%$ and p_2 is slightly more than this). So whatever the unknown state of Nature, this system of making assertions has an error rate at most 1.23%. With our data, we can assert H_2 with 98.77% confidence.

The same reasoning could be used for the conditionalist, with one modification. The maximum error rate is certainly 2.42% if H_1 is true, but it may differ from this if H_2 is true. In our case the appropriate calculation shows that if H_2 is true the maximum error rate is 2.09%. So our system of making statements still has a maximum error rate of 2.42%.

The general conclusion we would draw from this argument is that it is reasonable to make a *one-sided* test in favor of the hypothesis supported by the data, whenever the following hold:

(a) we are not sure of the direction of an effect;
but (b) we are sure it will not be exactly zero;
and (c) we have no clear null (status quo) hypothesis.

(Pratt comments that this "is utter heresy according to orthodox dogma.") Because of Pratt's argument, when we discuss Bayesian approaches which compute the posterior probability of H_1 after observing the data, we shall always compare these to the one-sided classical statements.

6. DISCUSSION OF THE BAYESIAN STATEMENTS

Suppose our prior density for p_1, p_2 is $f(p_1, p_2)$. After observing the data, the posterior probability that H_1 is true ($p_2 < p_1$) is

$$\frac{\int_{p_1=0}^{1}\int_{p_2=0}^{p_1} p_1^a(1-p_1)^b p_2^c(1-p_2)^d f(p_1, p_2)\,dp_2\,dp_1}{\int_{p_1=0}^{1}\int_{p_2=0}^{1} p_1^a(1-p_1)^b p_2^c(1-p_2)^d f(p_1, p_2)\,dp_2\,dp_1}.$$

If we take independent Haldane priors

$$f(p_1, p_2) \propto p_1^{-1}(1-p_1)^{-1}p_2^{-1}(1-p_2)^{-1},$$

the posterior probability becomes

$$\frac{1}{B(a,b)B(c,d)} \cdot \int_{x=0}^{1}\int_{y=0}^{x} x^{a-1}(1-x)^{b-1}y^{c-1}(1-y)^{d-1}\,dy\,dx$$

for $a, b, c, d > 0$, where $B(a, b)$ is the beta function. In the Appendix we call this probability $P(a, b, c, d)$ and derive some of its properties. One relation obtained shows that

$$P(a,b,c,d) - P(a-1,b+1,c+1,d-1) = H(a-1, a+b-1, a+c-1, a+b+c+d-2),$$

where $H(s, m, n, N)$ is (for consistent integer values of the parameters) the hypergeometric probability of obtaining s individuals having characteristics X and Y when m out of a total population of N are randomly assigned characteristic X and n out of the same population are randomly and independently assigned characteristic Y. Hence, for integer values of the parameters, P can be evaluated exactly as the tail of a hypergeometric distribution:

$$P(a,b,c,d) = \sum_{s<a} H(s, a+b-1, a+c-1, a+b+c+d-2).$$

Consequently it is precisely the p-value calculated by Fisher's exact test for the table $(a-1, b; c, d-1)$ when testing the null hypothesis $p_1 = p_2$ against the one-sided alternative H_2 ($p_1 < p_2$). This is all well known (see Altham, 1969). Altham also noted that if we started with the prior

$$f(p_1, p_2) \propto (1-p_1)^{-1}p_2^{-1},$$

our posterior for H_1 would be $P(a+1, b, c, d+1)$. This corresponds to Fisher's test on the table $(a, b; c, d)$ so Fisher's test *can* be given a Bayesian interpretation, but only if we start with an improper and unsymmetrical prior which favors H_1.

However, the exact test is well approximated by making Yates's correction when calculating the standard chi-square statistic. Hence we can approximate the exact test probability for the table $(a, b; c, d)$ by the lower tail of the standard normal distribution below z, where

$$z = \left\{\left(a+\frac{1}{2}\right)\left(d+\frac{1}{2}\right) - \left(b-\frac{1}{2}\right)\left(c-\frac{1}{2}\right)\right\} \cdot \sqrt{\frac{N}{m_1 m_2 n_1 n_2}}.$$

Combining these results, a good approximation to P should be given by

$$P(a, b, c, d) \approx \Phi(z),$$

where

$$z = \left\{\left(a-\frac{1}{2}\right)\left(d-\frac{1}{2}\right) - \left(b-\frac{1}{2}\right)\left(c-\frac{1}{2}\right)\right\} \cdot \sqrt{\frac{a+b+c+d-2}{(a+b-1)(c+d-1)(a+c-1)(b+d-1)}}.$$

(Altham discusses two approximations to P, one of which she describes as "the normal approximation to the hypergeometric distribution using Yates's half-correction," but in fact Yates's correction is not exactly the same as the continuity correction to the hypergeometric: there is a difference in the z-value by a factor $\sqrt{N/(N-1)}$.)

Starting with the Jeffreys prior,

$$f(p_1, p_2) \propto p_1^{-1/2}(1-p_1)^{-1/2}p_2^{-1/2}(1-p_2)^{-1/2},$$

the posterior probability of H_1 is given by $P(a + 1/2, b + 1/2, c + 1/2, d + 1/2)$. This can therefore be calculated approximately as $\Phi(z)$, where

$$z = \{ad - bc\}\sqrt{\frac{a+b+c+d}{(a+b)(c+d)(a+c)(b+d)}}.$$

But this is just the p-value for Yule's test of $p_1 = p_2$ against H_2 and corresponds to using the uncorrected chi-square (one-sided) test. So we have a Bayesian justification for the unconditional test!

The posterior probability can also be calculated exactly as the infinite sum of "hypergeometric probabilities":

$$P(a,b,c,d) = \sum_{-\infty<s<a} H(s, a+b-1, a+c-1, a+b+c+d-2).$$

However, now that a, b, c, d are not integers, some of the "probabilities" are negative. (H can be written in terms of gamma functions, and so can be evaluated for fractional arguments.) Alternatively, we can use a second equation from the Appendix,

$$P(a+1,b,c,d) - P(a,b,c,d) = \frac{1}{a}\frac{B(a+c, b+d)}{B(a,b)B(c,d)},$$

to calculate P as a finite sum (by starting from a base like $P(a, b; a, b)$ which is clearly equal to 1/2). Hence, for example, Barnard's 1947 example—the table $(0, 3; 3, 0)$—gives a posterior probability of $(1/2)-(5504/(1125\ \pi^2))$, or 0.43%. This can be compared to Yule's or Pearson's p-value of 0.72%. The exact test p-value is 5.00%, and the corrected chi-square statistic gives 5.12%.

For independent uniform priors we simply calculate $P(a+1, b+1, c+1, d+1)$ to find the posterior probability of H_1. Altham shows that, if the data favors H_2, this probability will be larger than that obtained when we use independent Jeffreys priors, which in turn will be larger than when independent Haldane priors are used.

However, should *independent* priors be used at all? We discuss this in the next section.

7. DEPENDENT PRIORS

Recall the example in the Introduction: do English or Scots cattle have a higher proportion of cows infected with a certain virus? Suppose we were informed (before collecting any data) that the proportion of English cows infected was 0.8. With independent uniform priors we would now give H_1 ($p_1 > p_2$) a probability of 0.8 (because the chance that $p_2 > 0.8$ is still 0.2). In very many cases this would not be appropriate. Often we will believe (for example) that if p_1 is 80%, p_2 will be near 80% as well and will be almost equally likely to be larger or smaller. (We are still assuming it will never be exactly the same.)

To formalize this idea we need dependent priors. Suppose that we think that when p is near zero, a doubling of p is important even though the absolute change in p is small, and similarly for changes in $1-p$ when p is near 1. Then we can express this by measuring p on a log-odds scale. Let

$$\theta_1 = \ln \frac{p_1}{1-p_1}, \quad \theta_2 = \ln \frac{p_2}{1-p_2}.$$

If θ_1 and θ_2 have independent improper uniform distributions on $(-\infty, \infty)$, this corresponds to Haldane improper priors. Suppose that, given θ_1, θ_2 has a normal distribution with mean θ_1 and standard deviation σ. This gives the sort of prior we require. It implies that knowledge of p_1 (or θ_1) would alter our probability distribution for θ_2 so that our expected value for θ_2 would then become the known value of θ_1. This "regression" model is in fact symmetrical between θ_1 and θ_2. The improper joint density function is proportional to

$$\exp\left\{-\frac{1}{2}\left(\frac{\theta_1-\theta_2}{\sigma}\right)^2\right\}.$$

In terms of p_1, p_2 this density is proportional to

$$e^{-(1/2)u^2} p_1^{-1}(1-p_1)^{-1} p_2^{-1}(1-p_2)^{-1},$$

where

$$u = \frac{1}{\sigma}\ln\left(\frac{p_1(1-p_2)}{p_2(1-p_1)}\right).$$

(Kass and Raftery, 1995, Section 7.1, describe a situation where it was possible to estimate σ from 12 related datasets.)

A generalization of the above would allow densities of the form proportional to

$$e^{-(1/2)u^2} p_1^{\alpha-1}(1-p_1)^{\beta-1} p_2^{\gamma-1}(1-p_2)^{\delta-1}$$

giving a large family of priors $S(\sigma, \alpha, \beta, \gamma, \delta)$. This family includes augmented versions of Haldane, Jeffreys and Laplace independent priors, the modification being made by the first term (which gives a sort of "closeness" factor and which creates the dependence). The posteriors also belong to this family (i.e., it is conjugate to the experiment), and it is clear that the closeness factor modifies the independent terms by shifting the bulk of the distribution towards the line $p_2 = p_1$, so that the posterior probabilities for H_1 and H_2 become less extreme. Figure 11 shows contour plots of augmented Jeffreys probability densities as functions of p_1 and p_2 on the unit square for $\sigma = 1/2$, 1 and 2, and the corresponding posterior densities after observing the Pearson data. For comparison, the independent Jeffreys prior density and its posterior are shown as well. The independent Jeffreys prior has a density which tends to infinity around the edges of the unit square, while the dependent prior densities are high along the diagonal $p_2 = p_1$ and tend to infinity at $(0, 0)$ and $(1, 1)$.

We recommend using a member of this family of prior densities for a Bayesian analysis of the 2×2 table. The smaller the value of σ (i.e., the more informative the prior), the more data is needed to give a large posterior probability for H_1 or H_2. In other words, in this problem the conservative thing to do is to choose an informative prior. Of course, a Bayesian should, as always, select a prior to reflect his own beliefs, but what should he use when reporting his results to others? We think that for this situation it is useful to have *standard* priors, and specifically for this problem we would suggest using the above dependent prior with $\sigma = 1$ and $\alpha = \beta = \gamma = \delta = 0$ as a standard prior. This is in fact improper, just as the Haldane prior itself is (see Lane and Sudderth, 1983), but it can be regarded as an approximation to the proper prior $S(1, \varepsilon, \varepsilon, \varepsilon, \varepsilon)$ for small $\varepsilon > 0$. Provided we observe at least one success and one failure, the posterior will be proper

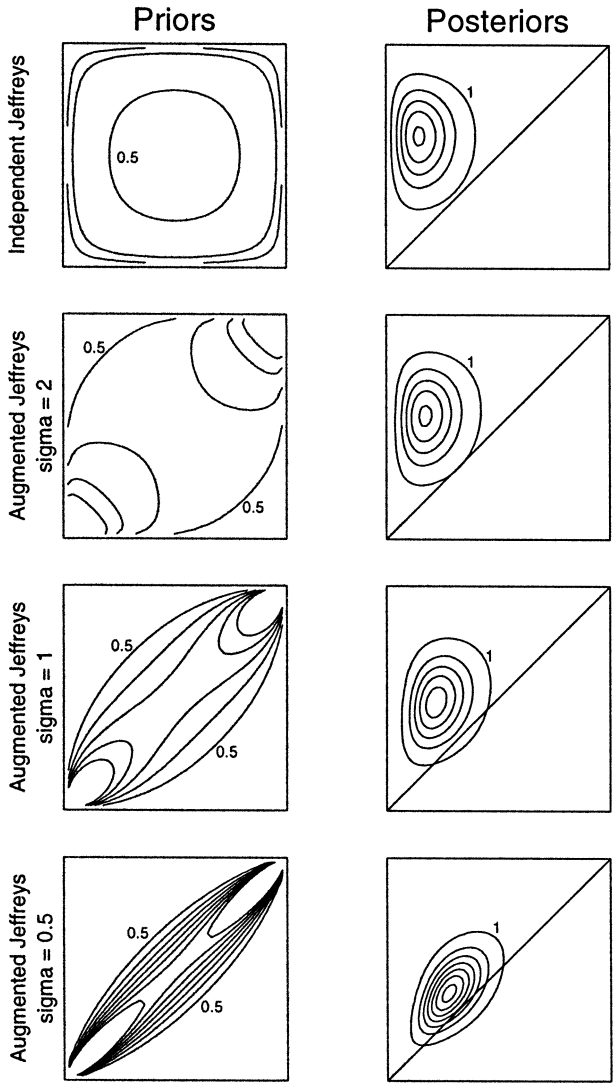

FIG. 11. *Probability densities on the unit square for independent Jeffreys priors and for augmented Jeffreys priors with* $\sigma = 2,\ 1$ *and* $1/2$: *the corresponding posteriors are also shown with the diagonal* $p_2 = p_1$. *For the priors, the lowest density contour is always* 0.5, *and the contours increase in steps of* 0.5. *For the posteriors, the lowest contour is* 1, *and the contours increase in steps of* 3.

and will be close to the posterior from the ε-prior. If we observed (say) no failures at all, we would need to think more carefully about our prior and select values for the five parameters of S to reflect our prior beliefs more precisely, because in this case the data is less helpful to us, and so we have to rely more heavily on these beliefs. In general, posteriors calculated from the standard prior could be reported as well as posteriors from the statistician's own prior. The next section gives some comparisons of the results of using the prior $S(\sigma, \alpha, \alpha, \alpha, \alpha)$ for $\alpha = 0,\ 1/2$ and 1 (augmented Haldane, Jeffreys and Laplace priors) and for $\sigma = 1/2,\ 1$ and 2.

With the standard prior, if we were told for certain that p_1 was 80%, we would then be 95% sure that p_2 was in the range 36–97%. If we knew p_1 was 50%, we would then be 95% sure that p_2 was in the range 12–88%. These do not seem very strong beliefs. If someone felt that intervals of this type were either far too wide or far too narrow, then the standard prior would not be appropriate even as a first approximation to his beliefs. He should then adjust σ to reflect his own prior more accurately.

Other approaches to modelling prior belief have been suggested in the literature. Antelman (1972) suggests using the Dirichlet-beta as a conjugate family of prior distributions. A hierarchical prior can also be used: for example, suppose that (conditional on μ, τ^2) θ_1 and θ_2 are IND(μ, τ^2) and that μ is assigned a vague prior. In fact, if the vague prior is taken as $N(\nu, \kappa^2)$, as $\kappa \to \infty$ we recover the prior $S(\sqrt{2}\tau, 0, 0, 0, 0)$. (However, if we also give τ a vague prior, we will lose the desired dependence.) If we wished to extend the ideas of this paper to more general situations where log-linear models are normally used, a hierarchical Bayesian model would be one natural way of doing this.

It is difficult to calculate these posterior probabilities analytically, but nowadays there is no problem in doing the integrations numerically. Our results were obtained using the *Mathematica* package. (It is perhaps worth noting that we will always obtain the same posterior probability for a table and for its transpose.) When we do this we get a final hypothetical statement about Pearson's data in Table 3.

THE BAYESIAN WITH DEPENDENT PRIOR. Suppose p_1 and p_2 were sampled from the standard dependent prior defined above. After the data was observed the posterior probability that $p_2 < p_1$ would be 2.99%.

Note that this statement is, in a sense, the most cautious we have seen. Even the conditionalist was quoting no more than 2.42%. Yet the prior did not seem that strong, so perhaps a good measure of caution is justified here.

8. COMPARISONS

For comparison (see Table 4), we give the various figures we have calculated for Pearson's table and a selection of others: (0, 3; 3, 0) is a famous example due to Barnard (1947); (5, 30; 11, 24) is from Berkson (1978); (2, 170; 9, 162) is a more recent example from Little (1989). The final table (20, 80; 30, 70) shows how the results converge with larger numbers in each cell.

TABLE 4
Comparisons of different calculations

Table	U	P	F	Y	L	J	H	S
(3, 15; 7, 5)	1.2	0.9	2.4	2.4	1.1	0.9	0.7	3.0
(2, 5; 5, 2)	9.0	5.4	14.3	14.3	6.6	5.3	4.0	13.1
(1, 4; 4, 1)	5.5	2.9	10.3	10.3	4.0	2.7	1.4	11.4
(0, 3; 3, 0)	1.6	0.7	5.0	5.1	1.4	0.4	—	9.5
(5, 30; 11, 24)	5.8	4.4	7.7	7.7	4.7	4.3	3.9	6.5
(2, 170; 9, 162)	1.6	1.6	3.0	3.2	1.7	1.3	1.0	2.9
(20, 80; 30, 70)	5.3	5.1	7.1	7.1	5.2	5.1	5.0	5.9

We show the table as $(a, b; c, d)$: U is the figure calculated by the unconditionalist; P is the uncorrected Pearson chi-square one-sided probability; F is from Fisher's exact test; Y is Yates's approximation to this; L, J and H are the Bayesian posteriors with Laplace, Jeffreys', and Haldane priors, respectively; and S is the recommended dependent prior.

The recommended prior tends to give more conservative results than the exact test for tables with very small entries, but is less conservative for more typical tables.

Table 5 shows the results for a range of dependent priors. It repeats the P and F columns from Table 4 and gives Bayesian posterior probabilities with priors $S(\sigma, \alpha, \alpha, \alpha, \alpha)$ for $\alpha = 0$, 1/2 and 1 and for $\sigma = 1/2$, 1 and 2; $\mathrm{AL}\frac{1}{2}$, for example, is augmented Laplace ($\alpha = 1$) with $\sigma = 1/2$, and similarly for the other columns (S is AH1). This table shows clearly that altering the strength of the dependence between the priors (the value for σ) can make very significant differences in the calculated posterior probability for H_1, whereas the choice between Laplace, Jeffreys, and Haldane in general makes only small changes in the result.

9. CONCLUSIONS

The main recommendation for Bayesians is to think carefully about using dependent priors. Even in the 2 × 2 table this can make a noticeable difference. Analyses of more general contingency tables often look at hypotheses in a large dimensional space (rather than the two dimensions we have been working in). We suspect that in these cases even a large amount of data will not always swamp the prior, and it may be important to consider whether independent priors are really justified.

The use of the uncorrected Pearson chi-square test for the 2 × 2 table corresponds approximately to a Bayesian analysis using independent Jeffreys priors. Because the priors are independent, we feel this test is not sufficiently cautious (and so agree with Fisher's conclusion, although for a different reason). The recommended alternative would be to use dependent priors, as in Section 7, but for statisticians who wish to use a classical approach, the best option would seem to be the exact test. Even that may not be cautious enough for small sample sizes.

Finally we reiterate that we have always excluded the case where $p_1 = p_2$ is a serious possibility. The problem of testing precise hypotheses is well discussed in Berger and Delampady (1987). Our analysis applies only to the case of imprecise hypotheses.

APPENDIX

Suppose we have independent Haldane (improper) priors. To calculate the posterior probability that $p_1 < p_2$ or $p_2 < p_1$ we need to evaluate integrals of the form

$$P(a, b, c, d) = \frac{1}{B(a, b)B(c, d)} \cdot \int_{x=0}^{1} \int_{y=0}^{x} x^{a-1}(1-x)^{b-1} y^{c-1}(1-y)^{d-1}\, dy\, dx$$

for $a, b, c, d > 0$, where $B(a, b)$ is the beta function

$$B(a, b) = \int_{x=0}^{1} x^{a-1}(1-x)^{b-1}\, dx.$$

It is easy to see that

$$P(a, b, c, d) = P(a, c, b, d) = P(d, c, b, a) = P(d, b, c, a);$$

TABLE 5
Comparisons of different dependent priors

Table	P	F	$\mathrm{AL}\frac{1}{2}$	AL1	AL2	$\mathrm{AJ}\frac{1}{2}$	AJ1	AJ2	$\mathrm{AH}\frac{1}{2}$	S	AH2
(3, 15; 7, 5)	0.9	2.4	10.3	3.4	1.6	10.2	3.2	1.4	10.2	3.0	1.2
(2, 5; 5, 2)	5.4	14.3	25.2	14.6	9.1	25.0	13.9	8.0	24.7	13.1	6.8
(1, 4; 4, 1)	2.9	10.3	24.7	13.1	6.8	24.4	12.3	5.6	24.1	11.4	4.3
(0, 3; 3, 0)	0.7	5.0	24.1	11.4	4.3	23.8	10.4	3.1	23.5	9.5	1.9
(5, 30; 11, 24)	4.4	7.7	13.4	7.3	5.4	13.1	6.9	5.0	12.8	6.5	4.5
(2, 170; 9, 162)	1.6	3.0	9.2	3.7	2.2	8.9	3.3	1.8	8.5	2.9	1.4
(20, 80; 30, 70)	5.1	7.1	8.7	6.2	5.5	8.6	6.0	5.3	8.5	5.9	5.2

$$1 - P(a,b,c,d) = P(c,d,a,b)$$
$$= P(b,a,d,c)$$
$$= P(c,a,d,b) = P(b,d,a,c).$$

Let $I_x(a,b)$ be the incomplete beta function,

$$I_x(a,b) = \frac{1}{B(a,b)} \int_{t=0}^{x} t^{a-1}(1-t)^{b-1}\,dt, \quad 0 \le x \le 1.$$

Then

$$P(a,b,c,d) = \frac{1}{B(a,b)} \int_{x=0}^{1} x^{a-1}(1-x)^{b-1} I_x(c,d)\,dx.$$

A recurrence equation can be derived by integrating by parts (see Abramowitz and Stegun, 1970):

$$I_x(c,d) = \frac{\Gamma(c+d)}{\Gamma(c+1)\Gamma(d)} x^c(1-x)^{d-1} + I_x(c+1,d-1).$$

Using this we obtain

$$P(a,b,c,d)$$
$$= \frac{\Gamma(c+d)}{\Gamma(c+1)\Gamma(d)} \frac{B(a+c,b+d-1)}{B(a,b)}$$
$$+ P(a,b,c+1,d-1)$$
$$= \frac{\Gamma(a+b)\Gamma(c+d)\Gamma(a+c)\Gamma(b+d-1)}{\Gamma(a)\Gamma(b)\Gamma(c+1)\Gamma(d)\Gamma(a+b+c+d-1)}$$
$$+ P(d-1,c+1,b,a).$$

Applying this result to $P(d-1,c+1,b,a)$ we have

$$P(a,b,c,d)$$
$$= P(d-1,c+1,b+1,a-1)$$
$$+ \frac{\Gamma(a+b)\Gamma(c+d)\Gamma(a+c)\Gamma(b+d-1)}{\Gamma(a)\Gamma(b)\Gamma(c+1)\Gamma(d)\Gamma(a+b+c+d-1)}$$
$$+ \frac{\Gamma(c+d)\Gamma(a+b)\Gamma(b+d-1)\Gamma(a+c)}{\Gamma(d-1)\Gamma(c+1)\Gamma(b+1)\Gamma(a)\Gamma(a+b+c+d-1)}.$$

Therefore

$$P(a,b,c,d) - P(a-1,b+1,c+1,d-1)$$
$$= \frac{\Gamma(a+b)\Gamma(c+d)\Gamma(a+c)\Gamma(b+d-1)(b+d-1)}{\Gamma(a)\Gamma(b+1)\Gamma(c+1)\Gamma(d)\Gamma(a+b+c+d-1)}$$
$$= \frac{\Gamma(a+b)\Gamma(c+d)\Gamma(a+c)\Gamma(b+d)}{\Gamma(a)\Gamma(b+1)\Gamma(c+1)\Gamma(d)\Gamma(a+b+c+d-1)}$$
$$= \frac{\binom{a+b-1}{a-1}\binom{c+d-1}{d-1}}{\binom{a+b+c+d-2}{a+c-1}}$$
$$= H(a-1,a+b-1,a+c-1,a+b+c+d-2),$$

where $H(s,m,n,N)$ is (for consistent integer values of the parameters) the hypergeometric probability defined in Section 6.

However, H is defined in terms of gamma functions for nonintegral values of the parameters. So we can express P as an infinite sum, as noted in Section 6. This sum generally converges (for $a, b, c, d > 0$), which suggests the domain of definition of P could be extended—perhaps to a, b, c and d having positive row and column sums, or a positive total sum. We shall not pursue this further here.

A second recurrence equation (substitute $\cos^2(\theta)$ for x) is

$$I_x(c,d) = \frac{\Gamma(c+d)}{\Gamma(c+1)\Gamma(d)} x^c(1-x)^d + I_x(c+1,d).$$

We can use this to obtain

$$P(a,b,c,d) = \frac{\Gamma(c+d)}{\Gamma(c+1)\Gamma(d)} \frac{B(a+c,b+d)}{B(a,b)}$$
$$+ P(a,b,c+1,d)$$
$$P(c,d,a,b) = P(c+1,d,a,b)$$
$$- \frac{\Gamma(c+d)}{\Gamma(c+1)\Gamma(d)} \frac{B(a+c,b+d)}{B(a,b)}.$$

Renaming variables we get

$$P(a+1,b,c,d) - P(a,b,c,d)$$
$$= \frac{\Gamma(a+b)}{\Gamma(a+1)\Gamma(b)} \frac{B(a+c,b+d)}{B(c,d)}$$
$$= \frac{1}{a} \frac{B(a+c,b+d)}{B(a,b)B(c,d)}$$
$$= \frac{1}{a} \frac{\Gamma(a+b)\Gamma(c+d)\Gamma(a+c)\Gamma(b+d)}{\Gamma(a)\Gamma(b)\Gamma(c)\Gamma(d)\Gamma(a+b+c+d)}.$$

Using this result we can calculate P for nonintegral parameters having the same fractional part as a finite sum, by starting from a base like $P(a,b,a,b)$ which is clearly equal to 1/2. In detail, define

$$g(a,b,c,d) = \frac{\Gamma(a+b)\Gamma(c+d)\Gamma(a+c)\Gamma(b+d)}{\Gamma(a)\Gamma(b)\Gamma(c)\Gamma(d)\Gamma(a+b+c+d)}.$$

Then we can calculate $P(a+s,b+s,c+s,d+s)$ for integer $a,b,c,d \ge 0$ and fractional s (e.g., $s = 1/2$ for Jeffreys priors) by starting from $P(1/2,1/2,1/2,1/2) = 1/2$ and adding a finite number of terms:

$$P(a+s,b+s,c+s,d+s)$$
$$= \frac{1}{2} + \sum_{i=0}^{a-1} \frac{g(i+s,s,s,s)}{i+s}$$

$$- \sum_{i=0}^{b-1} \frac{g(a+s, i+s, s, s)}{i+s}$$

$$- \sum_{i=0}^{c-1} \frac{g(a+s, b+s, i+s, s)}{i+s}$$

$$+ \sum_{i=0}^{d-1} \frac{g(a+s, b+s, c+s, i+s)}{i+s}.$$

Finally, we note that for general $a, b, c, d \geq 0$ we could calculate $P(a, b, c, d)$ by increasing the values of the four variables by integer amounts until we had a table for which P was approximately 1/2, or which could be very well approximated by a normal tail area.

ACKNOWLEDGMENTS

I thank the referees and Editors for many helpful comments and suggestions, which have led to major improvements in this paper.

REFERENCES

ABRAMOWITZ, M. and STEGUN, I., eds. (1970). *Handbook of Mathematical Functions*. National Bureau of Standards, U.S. Government Printing Office, Washington, DC.

ALTHAM, P. M. E. (1969). Exact Bayesian analysis of a 2×2 contingency table, and Fisher's "exact" significance test. *J. Roy. Statist. Soc. Ser. B* **31** 261–269.

ANTELMAN, G. R. (1972). Interrelated Bernoulli processes. *J. Amer. Statist. Assoc.* **67** 831–841.

BARNARD, G. A. (1947). Significance tests for 2×2 tables. *Biometrika* **34** 123–138.

BARTLETT, R. H., ROLOFF, D. W., CORNELL, R. G., ANDREWS, A. F., DILLON, P. W. and ZWISCHENBERGER, J. B. (1985). Extracorporeal circulation in neonatal respiratory failure: a prospective randomised study. *Pediatrics* **76** 479–487.

BEGG, C. B. (1990). On inferences from Wei's biased coin design for clinical trials. *Biometrika* **77** 467–484.

BERGER, J. O., BOUKAI, B. and WANG, Y. (1997). Unified frequentist and Bayesian testing of a precise hypothesis (with discussion). *Statist. Sci.* **12** 133–160.

BERGER, J. O. and DELAMPADY, M. (1987). Testing precise hypotheses. *Statist. Sci.* **2** 317–352.

BERKSON, J. (1978). In dispraise of the exact test. *J. Statist. Plann. Inference* **2** 27–42.

CHURCH, A. (1940). On the concept of a random sequence. *Bull. Amer. Math. Soc.* **46** 130–135.

CONOVER, W. J. (1974). Some reasons for not using the Yates continuity correction on 2×2 contingency tables (with discussion). *J. Amer. Statist. Assoc.* **69** 374–382.

CORNELL, R. G., LANDENBERGER, B. D. and BARTLETT, R. H. (1986). Randomized play the winner clinical trials. *Comm. Statist. Theory Methods* **15** 159–178.

CORNFIELD, J. (1956). A statistical problem arising from retrospective studies. *Proc. Third Berkeley Symp. Math. Statist. Probab.* 135–148. Univ. California Press, Berkeley.

COX, D. R. and HINKLEY, D. V. (1974). *Theoretical Statistics.* Chapman and Hall, London.

D'AGOSTINO, R. B., CHASE, W. and BELANGER, A. (1988). The appropriateness of some common procedures for testing the equality of two independent binomial proportions. *Amer. Statist.* **42** 198–202.

DAWID, A. P. (1984). Statistical theory: the prequential approach (with discussion). *J. Roy. Statist. Soc. Ser. A* **147** 278–292.

EDWARDS, A. W. F. (1972). *Likelihood.* Cambridge Univ. Press.

FISHER, R. A. (1935). *The Design of Experiments.* Oliver and Boyd, Edinburgh.

FISHER, R. A. (1945). A new test for 2×2 tables. *Nature* **156** 388.

FRANCK, W. E. (1986). *P*-values for discrete test statistics. *Biometrical J.* **28** 403–406.

FRASER, D. A. S., MONETTE, G. and NG, K.-W. (1984). Marginalization, likelihood and structural models. In *Multivariate Analysis VI* (P. R. Krishnaiah, ed.) 209–217. North-Holland, Amsterdam.

GOLDSTEIN, M. and HOWARD, J. V. (1991). A likelihood paradox. *J. Roy. Statist. Soc. Ser. B* **53** 619–628.

GRIZZLE, J. E. (1967). Continuity correction in the χ^2-test for 2×2 tables. *Amer. Statist.* **21** 28–32.

HABER, M. (1986). A modified exact test for 2×2 contingency tables. *Biometrical J.* **28** 455–463.

JEFFREYS, H. (1961). *Theory of Probability*, 3rd ed. Oxford Univ. Press.

KASS, R. E. and RAFTERY, A. E. (1995). Bayes factors. *J. Amer. Statist. Assoc.* **90** 773–795.

LANCASTER, H. O. (1961). Significance tests in discrete distributions. *J. Amer. Statist. Assoc.* **56** 233–234.

LANE, D. A. and SUDDERTH, W. D. (1983). Coherent and continuous inference. *Ann. Statist.* **11** 114–120.

LEHMANN, E. L. (1986). *Testing Statistical Hypotheses*, 2nd ed. Wiley, New York.

LITTLE, R. J. A. (1989). Testing the equality of two independent binomial proportions. *Amer. Statist.* **43** 283–288.

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.

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. *Phil. Mag.* **5**(50) 157–175.

PRATT, J. W. (1965). Bayesian interpretation of standard inference statements (with discussion). *J. Roy. Statist. Soc. Ser. B* **27** 169–203.

ROBBINS, H. (1977). A fundamental question of practical statistics. Letter to the editor. *Amer. Statist.* **31** 97.

ROUTLEDGE, R. D. (1992). Resolving the conflict over Fisher's exact test. *Canad. J. Statist.* **20** 201–209.

STONE, M. (1969). The role of significance testing: some data with a message. *Biometrika* **56** 485–493.

SUISSA, S. and SHUSTER, J. J. (1985). Exact unconditional sample sizes for the 2×2 binomial trial. *J. Roy. Statist. Soc. Ser. A* **148** 317–327.

TOCHER, K. D. (1950). Extension of Neyman–Pearson theory of tests to discontinuous variates. *Biometrika* **37** 130–144.

UPTON, G. J. G. (1992). Fisher's exact test. *J. Roy. Statist. Soc. Ser. A* **155** 395–402.

WEI, L. J. (1988). Exact two-sample permutation tests based on the randomized play-the-winner rule. *Biometrika* **75** 603–606.

WEI, L. J. and DURHAM, S. (1978). The randomized play-the-winner rule in medical trials. *J. Amer. Statist. Assoc.* **73** 830–843.

YATES, F. (1984). Tests of significance for 2×2 contingency tables (with discussion). *J. Roy. Statist. Soc. Ser. A* **147** 426–463.