JIRSS (2002)
Vol. 1, Nos. 1-2, pp 35-53

Another View of the Classical Problem of Comparing Two Probabilities

Herman Chernoff

Department of Statistics, Harvard University, Cambridge, MA 02138, USA. (chernoff@hustat.harvard.edu)

Abstract. The usual calculation of the P-value for the classical problem of comparing probabilities is not always accurate. This issue arose in the context of a legal dispute which depended on when some written material was written in a diary. The problem raises some issues on the foundations of statistical inference.

1 Background

A legal case concerned the question of when a certain handwritten entry was made in a diary. One party contended that two paragraphs on a page were written before the entries on the reverse side of that page. It was admitted that the writing was all made under similar conditions by one writer. The other party claimed that the short first paragraph was written after the writing on the reverse side of that paragraph.

The writing was subjected to forensic examination by Erich Speckin of Speckin Forensic Laboratories. The underlying theory behind his

Received: May 2002, Revised: September 2002

Key words and phrases: Ancillary, Bayes, binomial, Fisher exact test, likelihood-ratio, p-values, 2×2 table, Yates continuity correction.

examination is that the earlier writing leaves a convex impression on the page where the pen moves. The later writing on the reverse side introduces a number of *intersections* where the ink trails from the two sides cross. The convexity of the first impression often causes the later impression on the reverse side to leave a short *skipping*, a gap in the ink trail, which is frequently detectable by examination of enlarged photographs, when the intersection is at an appropriate angle.

The examination is not error free. Thus an intersection where a skipping should be detected does not always yield an observable skipping, whereas an intersection where a skipping should not be observed might lead to what the forensic examiner regards as an observable skipping. If the two error probabilities are ϵ_f and ϵ_l for the first and later sections written, then the probabilities of observing a skipping at an intersection are $q_f = \epsilon_f$ and $q_l = 1 - \epsilon_l$ respectively. If the two error probabilities are small enough, q_l is substantially greater than q_f. Then the observed skippings at a reasonable number of intersections should provide evidence to determine whether a paragraph was written earlier or later than the writing on the reverse side.

As we shall see in the following section, the issue involves a one-sided test of the hypothesis that two unknown probabilities are equal. This, in turn, raises the question of how to calculate a *conservative* version of the P-value for this test and the inferential meaning of such a calculation.

2 Several relevant hypotheses

Both parties are willing to grant that the writings on both sides of the page were made by one person under similar conditions. One party claims that the two paragraphs on the first side were written before the material on the reverse side. The contending party claims that the first short paragraph on the first side was written later than the material on the reverse side. It accepts the claim that the second paragraph on the first side was written earlier than the material with which it intersects on the reverse side.

To test the reliability of the claim of the first party we should test the hypothesis $H_1 : p_1 = p_2 = q_f$ vs. $K_1 : p_1 = q_l > p_2 = q_f$, where p_1 and p_2 are the probabilities of observing a skipping for the first and second paragraphs. Mr. Speckin had observed $X_1 = 7$ skippings out of $n_1 = 16$ intersections for the first paragraph, and $X_2 = 24$ skippings out of $n_2 = 261$ intersections for the second paragraph. These data lead to the estimates $\hat{p}_1 = 7/16 = 0.4375$ and $\hat{p}_2 = 24/261 = 0.092$ for p_1 and p_2. From an informal point of view it appears unlikely that the probability of observing a skipping at an intersection is the same for both paragraphs.

The skippings due to the writing on the reverse side provide additional relevant data. It seemed to me that these data should not be ignored. Indeed, the legal case for claiming a difference between p_1 and p_2 might be weakened if the judge were suspicious about why these data were not presented. At my request, Speckin examined skippings due to writing on the reverse side. He had previously found $X'_1 = 0$ opposite the first paragraph, and then he obtained $X'_2 = 126$ opposite the second paragraph.

Considering the new data, we are led to introduce additional hypotheses. These are $H_2 : p'_1 = p'_2$, $H_3 : p_1 = p_2 < p'_1 = p'_2$, $K_2 : p'_2 > p'_1$, and $K_3 : p_1 = p'_2 > p_2 = p'_1$ where p'_1 and p'_2 are the probabilities of an observed skipping for the reverse side of paragraphs 1 and 2 respectively.

According to the claim that the material on the first side was written before that of the reverse side, we should expect to find that the data are consistent with H_1, H_2, and H_3 and we should test these against the corresponding alternatives K_1, K_2, and K_3.

Since q_f and q_l are unknown characteristics of the writer, the parameters p_1, p_2, p'_1 and p'_2 are unknown.

3 Three standard one-sided tests for the equality of two probabilities

The problem of testing H_1 vs. K_1 is one of the most studied problems in the statistical literature. When I declared to two colleagues that I had a new insight on the problem, each of them referred me to relatively recent papers that they had published on this subject Little (1989) and Rubin and Stern (1994).

For most purposes, the standard approaches are pretty reasonable. In my case, a superficial view of the data is compelling. It was my desire to be more conservative than necessary that led me to look further. I shall list three of the standard approaches first.

One of the standard approaches, which we will call the Pearson method, is to use as the test statistic

$$Z_P = \frac{X_1/n_1 - X_2/n_2}{\sqrt{\hat{p}(1-\hat{p})(n_1^{-1}+n_2^{-1})}} \tag{1}$$

where $n = n_1 + n_2$ and $\hat{p} = (X_1 + X_2)/n$ is the maximum likelihood estimate of the common value of p_1 and p_2 under the hypothesis H_1. If $p_1 > p_2$, Z_P will tend to be positive and the one-tail test of H_1 consists of rejecting H_1 if Z_P is sufficiently large. Under H_1 the asympototic distribution of Z_P is $N(0,1)$, the normal distribution with mean 0 and variance 1. The P-value for this test is taken to be $P_P = p_P(Z_P)$ where

$$p_P(z) = P(Z_P \geq z|H_1) = 1 - \Phi(z) \tag{2}$$

and Φ is the cumulative distribution function for the standard normal distribution $N(0,1)$.

In the early history, lack of sufficient computer power made it important to develop an asymptotic theory, as a result of which some of the claims are approximations which become poor when the sample sizes are not large. Some of the recent publications, Little (1989), seek to determine how conservative some of these approximations are. In our application, n_2 is reasonably large, but n_1 is quite moderate in size.

A second approach, which we shall label the Yates approach, applies a *continuity correction* which improves the asymptotic approximation. This method can be explained by constructing a 2 × 2 table with entries $a = X_1$, and $b = X_2$ in the first row, and $c = n_1 - X_1$ and $d = n_2 - X_2$ in the second row. The table has margins n_1 and n_2 at the bottom and $m_1 = X_1 + X_2$ and $m_2 = n - m_1$ on the side. See Table 1.

Table 1: 2 × 2 table

$a = X_1$	$b = X_2$	$m_1 = X_1 + X_2$
$c = n_1 - X_1$	$d = n_2 - X_2$	$m_2 = n - m_1$
n_1	n_2	n

The Yates test statistic is

$$Z_Y = \sqrt{\frac{n}{n_1 n_2 m_1 m_2}}(ad - bc - n/2) \tag{3}$$

which corresponds to the chi-square statistic and Z_P, except that a and d are decreased by $1/2$ and b and c are increased by $1/2$. Here, the P-value is given by $P_Y = 1 - \Phi(Z_Y)$.

A third approach is that of the Fisher Exact Test. This test capitalizes on the fact that the conditional probability of observing the 2×2 table, given the margins m_1, m_2, n_1 and n_2, does not depend on the unknown common value of p_1 and p_2 under the hypothesis. This conditional probability is

$$z_F(a) = n_1!n_2!m_1!m_2!/n!a!b!c!d! \tag{4}$$

Note that for specified values of the margins the value of a determines all the other entries. Thus z_F may be regarded as a function of a, and we may treat $Z_F = z_F(X_1)$ as the test statistic (conditional on $X_1 + X_2$), with conditional P-value

$$P_F = \sum_{a=X_1}^{a^*} z_F(a) \tag{5}$$

where $a^* = \min(n_1, m_1)$ and $m_1 = X_1 + X_2$.

For the data given, the P-values for the three approaches on the test of H_1 vs. K_1 are $P_P = 1.0(-5), P_Y = 6.0(-5)$, and $P_F = 6.3(-4)$. In these representations the integer in parentheses represents the exponent of 10. These results bear out the general claim that both P_F and P_Y tend to be more conservative than P_P in the sense that they do not argue as strongly as P_P against the hypothesis being tested. The fact that P_Y is so much greater than P_P suggests that the asymptotic approximation needs the continuity correction in this case where the data strongly suggest that the hypothesis is false and at least one of the sample sizes is not large.

With the use of modern computers it is no longer necessary to rely on asymptotic approximations, and while P_F is exact, both of the other P-values can be also computed exactly, assuming that the common value of p under H_1 is $\hat{p}$, to yield $P_{P1} = 2.9(-4)$ and $P_{Y1} = 2.9(-4)$. The relatively large discrepancy between the exact and asymptotic results is apparently due, in part, to the large deviation effect in the tails of the distribution of the test statistic, as well as to the modest sample size n_1.

However, we have a more serious problem due to the fact that the hypothesis H_1 is a composite hypothesis, and the distribution of the data depend on the unknown common value of p_1 and p_2 under the hypothesis. That problem will be attacked in the next section.

4 The effect of composite hypotheses

Technically speaking, the P-value for a composite hypothesis is defined differently than what we have apparently done. If we have a composite hypothesis $H : \theta \in \Theta$ and a test statistic T, then the P-value is $P = \sup_{\theta \in \Theta} p(T; \theta)$ where $p(t; \theta) = P(T \geq t | H, \theta)$. In using P_P and P_Y, we implicitly assumed that the asymptotic normal distribution $N(0, 1)$, which is independent of the *nuisance parameter*, the common value of p_1 and p_2 assumed under H_1, represented the true distribution of the test statistic. This asympotic distribution is not a very good fit in the tails; the true distribution of Z_P and Z_Y depends on p, and a true P-value should take into account $P(p) = p(T; p)$ for all values of $p \in (0, 1)$.

Although the Fisher approach provides an exact test, that test is exact only if we condition on the margins. That procedure is appropriate in some tests of independence for 2×2 tables, but it is not exactly so here. In the case of testing for independence with specified margins, $X_1 + X_2$ would be an *ancillary* statistic, but it is only approximately ancillary in our problem Little (1989). Thus even the Fisher test yields a range of P-values depending on the nuisance parameter, and we should consider the maximum of these P-values. Note that to define a P-value, one must have a test statistic. For the case where the marginals are given, we could use X_1 as the test statistic. But if the marginals are not specified, a more natural test statistic would be the usual P-value using the Fisher exact test, assuming, without proper justification, that the marginals are specified.

Unfortunately, the range of P-values corresponding to various values of p is uncomfortably large for P_P and P_Y. Our calculations indicate that $P_P(p), P_Y(p)$ and $P_F(p)$ attain their maximum values of 3.29(-3), 1.03(-3), and 2.93(-4), at $p = 0.012, 0.0092$ and 0.15 respectively. On the other hand, if we do exact calculations, assuming that $\hat{p}$ is the only value of p to consider, we would have values of 2.91(-4), 2.89(-4), and 2.86(-4) respectively. These compare with our original values of 1.0(-5), 6.0(-5) and 6.3(-4) respectively.

One of our difficulties comes from the poor quality of the asymptotic approximations in this example where one of the sizes is moderate, and large deviation effects are relevant. Another major problem comes from the fact that using the proper definition of the P-value provides another large effect. The range of values from 1.0(-5) to 3.3(-3) is enormous, although not enough to make one believe in H_1. Note that $P_F(p)$ is very stable. Somewhat surprisingly $P_F(p)$ tends to be about 2.9(-4) over a large range of values of p, but different from $P_F = 6.3(-4)$. Figure 1 shows how these P-values depend on p.

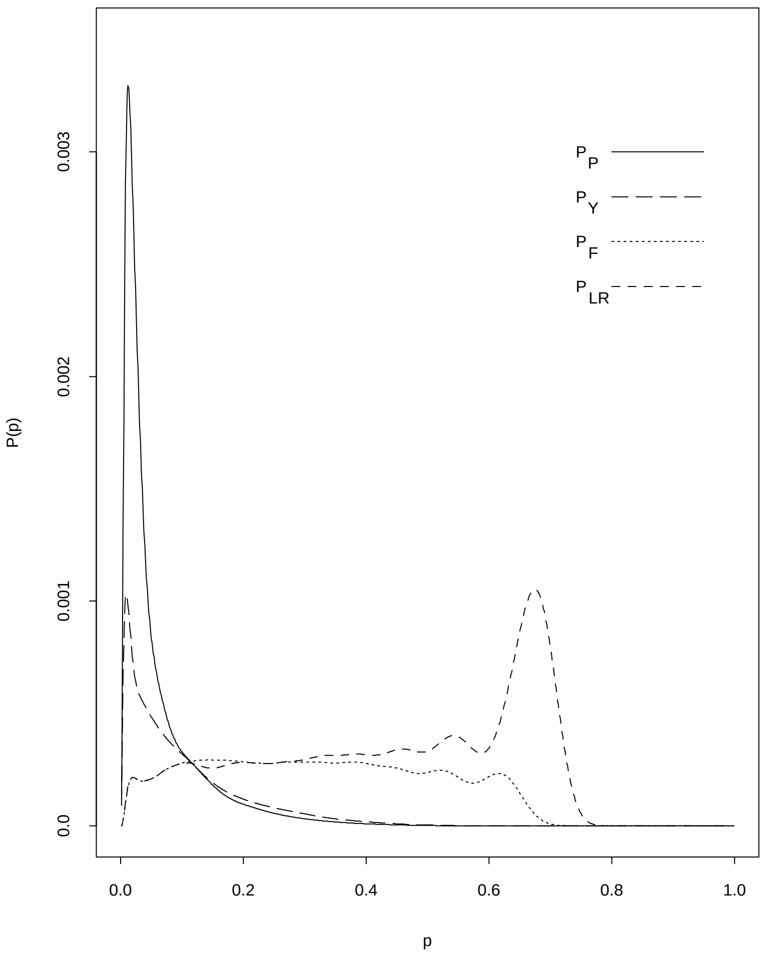

Figure 1: P-values for Pearson, Yates, Fisher, and LR

5 Modified P-values

While P-values have played a useful role in statistical inference, their use is subject to considerable criticism by Bayesians. One major limitation in the general case is that the P-value does not put enough emphasis on the role of the alternative hypotheses. What good is it to use the data to reject a hypothesis as unreasonable, if the same data would show that the potential alternative is also unreasonable? This latter criticism does not apply in the current problem. Granted that P-values have limited inferential value, does it pay to use a precise definition in this example? Here the maximum value of $P_P(p)$ is attained at $p = 0.012$ which is very inconsistent with the estimate $\hat{p} = 37/277 = 0.111$ assuming H_1. A 95% confidence interval for p would be (0.077,0.155). Over this range the maximum of the three measures $P_P(p), P_Y(p)$, and $P_F(p)$ are 4.5(-4), 3.8(-4), and 2.9(-4) respectively.

However, there is a mismatch between the confidence and P-values. It seems strange to say that "I am 95% confident that the P-value is 4.6(-4)". With my remaining confidence, the P-value could conceivably be be very large. It would seem more sensible to say something like "I am 99.9% confident that the P-value is 0.001" Increasing the confidence leads to increasing the range of potential values of p, and consequently increasing the maximum P-value over that range. With such matching we achieve approximately 99.95% confidence for $P_P(p) \leq 7.1(-4), P_Y(p) \leq 4.6(-4)$ and $P_F(p) \leq 2.9(-4)$. Thus a substantial increase in confidence leads to a relatively small increase in P-values. Table 2 presents various level confidence intervals of confidence γ for p, based on x successes out of n trials. Under H_1 the intervals for $(x, n) = (31, 277)$ are the confidence intervals for p while $(x, n) = (126, 277)$ is relevant for H_2.

Table 2. Intervals with confidence γ for a probability based on x successes out of n trials.

x/n	24/277		31/277		126/277		133/277	
1-γ								
5(-2)	0.05630	0.12617	0.07732	0.15508	0.39519	0.51553	0.41999	0.54072
1(-2)	0.04872	0.13940	0.06836	0.16933	0.37735	0.53398	0.40193	0.55906
5(-3)	0.04606	0.14454	0.06518	0.17484	0.37071	0.54089	0.39519	0.56591
1(-3)	0.04082	0.15562	0.05884	0.18666	0.35695	0.55529	0.38120	0.58018
5(-4)	0.03888	0.16009	0.05646	0.19141	0.35158	0.56093	0.37574	0.58577
1(-4)	0.03490	0.16995	0.05156	0.20183	0.34010	0.57305	0.36405	0.59774
5(-5)	0.03339	0.17399	0.04967	0.20610	0.33552	0.57797	0.35937	0.60254
1(-5)	0.03023	0.18302	0.04570	0.21560	0.32556	0.58852	0.34920	0.61300
5(-6)	0.02900	0.18677	0.04414	0.21953	0.32152	0.59283	0.34507	0.61725
1(-6)	0.02641	0.19520	0.04083	0.22834	0.31266	0.60234	0.33600	0.62660
5(-7)	0.02539	0.19872	0.03951	0.23202	0.30904	0.60624	0.33228	0.63044
1(-7)	0.02322	0.20668	0.03669	0.24030	0.30102	0.61490	0.32406	0.63894
5(-8)	0.02236	0.21002	0.03556	0.24377	0.29772	0.61848	0.32067	0.64245
1(-8)	0.02052	0.21759	0.03312	0.25162	0.29037	0.62646	0.31313	0.65028
5(-9)	0.01979	0.22078	0.03214	0.25493	0.28734	0.62977	0.31000	0.65352
1(-9)	0.01821	0.22803	0.03000	0.26242	0.28055	0.63719	0.30302	0.66079

The comparison of these three approaches seems to favor the Fisher approach, where the maximum P-value of $P_F(p) = 2.93(-4)$ is relatively small, and takes place at $p = 0.145$ which is not far from $\hat{p}$. The large values of $P_P(p)$ and $P_Y(p)$ for small values of p reflects the fact that the denominator of Z_P, which is a normalizing factor when p is close to $\hat{p}$, no longer serves that purpose well when p is very small.

6 Other approaches

Given the stability of the Fisher approach, it seems unlikely that there will be much room for improvement in this special problem of testing for the equality of two probabilities. For more complicated problems, *e.g.* testing H_3 vs. K_3, we may not have the benefit of the advice of Fisher to select such an effective test statistic. We propose to introduce two new methods, one using a Bayesian approach and both depending on the likelihood-ratio test statistic. We would hope that the use of the likelihood-ratio would free us of the need to depend on some clever choice of the test statistic.

We elaborate on the likelihood-ratio approach for which we use the abbreviation LR. The likelihood for this model is $L(p_1, p_2) = l(X_1, X_2; p_1, p_2)$ where

$$l(x_1, x_2; p_1, p_2) = p_1^{x_1}(1-p_1)^{n_1-x_1}p_2^{x_2}(1-p_2)^{n_2-x_2}.$$

Let Z_L be the log of the likelihood-ratio for the two hypotheses H_1 and K_1 calculated at the maximum likelihood estimates under the two hypotheses.

$$Z_L = ((n_1+n_2)H(\hat{p}) - n_1 H(\hat{p}_1) - n_2 H(\hat{p}_2))(\mathrm{sgn}(\hat{\mathrm{p}}_1 - \hat{\mathrm{p}}_2)) \tag{6}$$

where H is the entropy given by

$$H(p) = -[\mathrm{p}\log(p) + (1-p)\log(1-p)] \tag{7}$$

and $\hat{p}_1 = X_1/n_1$, $\hat{p}_2 = X_2/n_2$, and $\hat{p} = (X_1+X_2)/(n_1+n_2)$.

Then $p_L(z;p) = P(Z_L \geq z|H_1, p)$, $P_L(p) = p_L(Z_L;p)$, $P_{L1} = P_L(\hat{p})$, and

$$P_L(p) = \sum_{\mathbf{x}^* \in A} \binom{n_1}{x_1^*}\binom{n_2}{x_2^*} p^{x_1^*+x_2^*}(1-p)^{n_1+n_2-x_1^*-x_2^*} \tag{8}$$

where $\mathbf{x}^* = (x_1^*, x_2^*)$ and $A = \{\mathbf{x}^* : Z_L^* \geq Z_L\}$.

Finally we introduce P_B which is the posterior probability of the hypothesis H_1 based on a somewhat artificial and arbitrary prior distribution, for which the theoretical calculations are relatively simple. We assume that the prior distribution of p_1 and p_2 is a mixture of two distributions. One of these distributions is a beta distribution in p_0 on the line where $p_1 = p_2 = p_0$. The other distribution takes p_1 and p_2 to have independent beta distributions. With a slight abuse of notation we may write, for the distribution law $\mathcal{L}$ of (p_1, p_2),

$$\mathcal{L}(p_1, p_2) = wBe(a_0, b_0) + (1-w)Be(a_1, b_1) * Be(a_2, b_2) \tag{9}$$

to indicate that with probability w, $p_1 = p_2 = p_0$ which has a beta distribution with parameters (a_0, b_0) and with probability $1-w$, p_1 has the beta distribution with parameters (a_1, b_1) and p_2 has the beta distribution with parameters (a_2, b_2) and p_1 and p_2 are independent.

Then the posterior distribution, given the data, $\mathbf{X} = (X_1, X_2)$, has the same form and can be written

$$\mathcal{L}(p_1, p_2|\mathbf{X}) = w^*Be(a_0^*, b_0^*) + (1-w^*)Be(a_1^*, b_1^*) * Be(a_2^*, b_2^*) \tag{10}$$

where $P_B = w^*$ is the posterior probability of the hypothesis H_1. The necessary calculations are described as follows. If the data yield X_1 successes out of n_1 trials with probability p_1 and X_2 successes out of n_2 trials with probability p_2, then $a_1^* = a_1 + X_1$, $b_1^* = b_1 + n_1 - X_1$, $a_2^* = a_2 + X_2$, $b_2^* = b_2 + n_2 - X_2$, $a_0^* = a_0 + X_1 + X_2$, $b_0^* = b_0 + n_1 + n_2 - X_1 - X_2$, and

$$\frac{w^*}{1-w^*} = \frac{w}{1-w} \frac{B_1 B_2 B_0^*}{B_1^* B_2^* B_0} \tag{11}$$

where $B_i = B(a_i, b_i)$, $B_i^* = B(a_i^*, b_i^*)$, B is the beta function

$$B(c_1, c_2) = \Gamma(c_1)\Gamma(c_2)/\Gamma(c_1 + c_2)$$

and Γ is the gamma function.

Low values of a_i and b_i correspond to a prior beta distribution which, in some sense, assumes very little about the distribution of p_i. For example $a_1 = b_1 = 1$ yields a uniform distribution on the interval from 0 to 1 for p_1. On the other hand $a_1 = b_1 = 20$ would correspond to a prior distribution which assumes that p_1 tends to be rather close to 0.5. There is a good deal of subjectivity in the process of applying the Bayesian philosophy in our current problem, and I recommend it with reservations.

At this point let us summarize our results and notation on P-values. In general we have a test statistic T for testing a hypothesis H, against an alternative K, a conventional P-value, and a nuisance parameter $\theta \in \Theta$. Let $p(t;\theta) = P(T \geq t|H,\theta)$, $P_1 = p(T;\hat{\theta})$, where $\hat{\theta}$ is the maximum likelihood estimate of θ under H, and $P_2 = sup_{\theta \in \Theta} p(T;\theta)$. Given a 95% confidence interval P_3 is the supremum of $p(T;\theta)$ for θ in that interval. Similarly, P_4 corresponds to the maximum P-value over an approximately *matching* confidence interval Table 3 summarizes the various P-values for testing H_1. (The last two digits of the tabled values of P_4 are not reliable).

Table 3. P-values for testing H_1 vs. K_1.

	Pearson	Yates	Fisher	LR
P	1.04(-5)	5.97(-5)	6.28(-4)	
P_1	2.91(-4)	2.89(-4)	2.86(-4)	2.79(-4)
P_2	3.29(-3)	1.03(-3)	2.93(-4)	1.04(-3)
P_3	4.62(-4)	3.80(-4)	2.93(-4)	2.80(-4)
P_4	7.09(-4)	4.56(-4)	2.93(-4)	2.80(-4)

The posterior probability of H_1 for the parameters $w = 0.5, a_0 = b_0 = a_1 = b_1 = a_2 = b_2 = 1$ is $w^* = 8.94(-3)$. With this prior distribution, for which $P(p_1 < p_2) = 0.25$, there is a small posterior probability, 4.59(-4), that $p_1 < p_2$. When using a Bayesian approach it often makes sense to use information based on previous experience. Suppose that we started with $a_0 = b_0 = a_1 = b_1 = 5$ and $a_2 = 2, b_2 = 18$. I chose these numbers without any real background experience, and without reference to Mr. Speckin who may have some relevant suggestions, merely to represent a possibility. The $Be(5,5)$ and $Be(2,18)$ distributions have means 0.5 and 0.1, and standard deviations 0.151 and 0.065. With these parameters to represent the prior distribution, we obtain the posterior probabilty of H_1 to be $w^* = 5.76(-5)$.

7 The reverse side

Up to now, we have concentrated almost exclusively on the observed skippings on one side of the page. We have additional data from the reverse side where $X_1' = 0$ and X_2'=126. These additional data could be used to test H_2 vs. K_2 with the same methods used to test H_1 vs. K_1. Corresponding P-values are listed in Table 4.

Table 4. P-values for testing H_2 vs. K_2.

	Pearson	Yates	Fisher	LR
P	8.35(-5)	2.28(-4)	4.15(-5)	
P_1	1.34(-5)	1.34(-5)	1.22(-5)	1.21(-5)
P_2	2.29(-2)	1.21(-3)	1.67(-5)	1.42(-5)
P_3	3.64(-5)	3.64(-5)	1.44(-5)	1.42(-5)
P_4	7.28(-5)	7.28(-5)	1.47(-5)	1.42(-5)

The posterior probability of H_2 for the parameters $w = 0.5, a_0 = b_0 = a_1 = b_1 = a_2 = b_2 = 1$ is $w^* = 6.65(-5)$. The posterior probability that $p_1' > p_2'$ is 2.13(-5).

7.1 Combining P-values

One standard approach for combining the results of testing the two hypotheses H_1 and H_2 uses the fact that under the hypothesis to be tested the P-value has a uniform distribution and its negative logarithm has an exponential distribution, which may also be regarded as the gamma distribution with scale parameter 1 and shape parameter 1, *i.e.* $\Gamma(1, 1)$. Then the sum of two such independent random variables has the $\Gamma(1, 2)$ distribution, the tail probabilities of which can be used to provide a combined P-value. For example if the two P-values are 2.80(-4) and 1.42(-5) the sum of the two negative logarithms is 19.343 for which the tail probability is $P = 8.09(-8)$.

There are been three implicit assumptions in the analysis above. These are that the P-values are independent, that they have continuous distributions, and that the hypothesis being tested is *simple, i.e.* not composite. These assumptions are not satisfied here, but that probably has little influence here. The data are discrete and the P-values are not quite continuously distributed. The observed skippings, at an intersection on both sides of the page are probably not quite independent for all the candidate intersections. I believe that for a given value of $p, P(p)$ is reasonably well approximated by the combining anlaysis in spite of the failure of the assumptions. On the other hand this may not hold so well if we try to combine the maxima in this way.

7.2 Likelihood-ratio

An alternative to combining the P-values of the two tests is to test H_3 vs. K_3. Such a test introduces an interesting variation on our problem. It made little sense to get P-values for testing K_1 vs. H_1 or K_2 vs. H_2. As long as the observed proportions favored K_1 and K_2, such P-values could not be very small. One could not show that both H_1 and K_1 were inconsistent with the data. That is no longer the case for testing H_3 vs. K_3. Both of these alternative hypotheses are specific enough that each can be tested to see how consistent they are with the data. Is it the case that the skipping counts are not only inconsistent with H_3, but also inconsistent with the assumptions that (i) the first paragraph and the reverse of the second paragraph show similar probabilities of observed skippings, (ii) the same can be said for the second paragraph and the reverse of the first, and (iii) the probability for the first pair above exceeds that of the second pair?

The likelihood-ratio and Bayesian approaches are applicable for testing H_3 vs. K_3. These methods may be applied in a routine fashion and do not require the use of deep insight to develop a clever test statistic. On the other hand the resulting method does not directly confront assumptions (i) and (ii). To do so requires a modified test.

In this subsection we shall describe the likelihood-ratio method. We introduce the variables $R_{12} = (X_1 + X_2)/n, R_{34} = (X_1' + X_2')/n$, $R_{14} = (X_1+X_2')/n, R_{23} = (X_2+X_1')/n, R_0 = (X_1+X_2+X_1'+X_2')/2n$, and $\mathbf{X} = (X_1, X_2, X_1', X_2')$. Under H_3, the maximum of the logarithm of the likelihood is given by

$$\begin{array}{rcll} l(\mathbf{X}|H_3) & = & -n[H(R_{12}) + H(R_{34})] & \text{if } R_{12} < R_{34} \\ & = & -2nH(R_0) & \text{if } R_{12} \geq R_{34} \end{array}$$

Similarly, the logarithm of the likelihood under K_3 attains its maximum of

$$\begin{array}{rcll} l(\mathbf{X}|K_3) & = & -n[H(R_{23}) + H(R_{14})] & \text{if } R_{23} < R_{14} \\ & = & -2nH(R_0) & \text{if } R_{23} \geq R_{14} \end{array}$$

and we may use

$$Z_T = l(\mathbf{X}|K_3) - l(\mathbf{X}|H_3) \tag{12}$$

as our test statistic.

The P-value for testing H_3 is $P_H = p_H(Z_T)$ where

$$p_H(z) = sup_{p_1<p_1'} P(Z_T \geq z|H_3, p_1, p_1') \tag{13}$$

Similarly, the P-value for testing K_3 is $P_K = p_K(Z_T)$ where

$$p_K(z) = sup_{p_1 > p_1'} P(Z_T \leq z | K_3, p_1, p_1') \tag{14}$$

Table 5 presents the results for testing H_3 and K_3 in analogy with the notation of Table 3.

Table 5. P-values for testing H_3 vs. K_3 and K_3 vs. H_3

	P_H	P_K
P		
P_1	3.33(-8)	6.02(-1)
P_2	5.05(-8)	1.00(-0)
P_3	3.49(-8)	9.73(-1)
P_4	5.91(-5)	9.73(-1)

7.3 Bayesian approach

If we assign prior probabilities w_3 and $1 - w_3$ to H_3 and K_3, a natural version of the prior distribution of (p_1, p_1') may be written as

$$\mathcal{L}(p_1, p_1') = \\ w_3 Be(a_1, b_1) * Be(a_2, b_2) + (1 - w_3) Be(a_2, b_2) * Be(a_1, b) (15)$$

Letting $\mathbf{X}$ represent the complete data, we have

$$\mathcal{L}(\mathbf{X} | H_3, p_1, p_1') = \\ p_1^{X_1 + X_2} (1 - p_1)^{n - X_1 - X_2} (p_1')^{X_1' + X_2'} (1 - p_1')^{n - X_1' - X_2'} \tag{16}$$

and

$$\mathcal{L}(\mathbf{X} | K_3, p_1, p_1') = \\ p_1^{(X_1 + X_2')} (1 - p_1)^{(n - X_1 - X_2')} (p_1')^{(X_1' + X_2)} (1 - p_1')^{(n - X_1' - X_2)} . (17)$$

It follows that the posterior distribution is given by

$$\mathcal{L}(p_1, p_1' | \mathbf{X}) = \\ w_3^* B_1 * B_2 + (1 - w_3^*) B_3 * B_4 \tag{18}$$

where $B_i = Be(a_i^*, b_i^*)$ for $1 \leq i \leq 4$, $a_1^* = a_1 + X_1 + X_2$, $b_1^* = b_1 + n - X_1 - X_2$, $a_2^* = a_2 + X_1' + X_2'$, $b_2^* = b_2 + n - X_1' - X_2'$, $a_3^* =$

$a_2 + X_1 + X_2'$, $b_3^* = b_2 + n - X_1 - X_2'$, $a_4^* = a_1 + X_1' + X_2$, $b_4^* = b_1 + n - X_1' - X_2$, and

$$\frac{w_3^*}{1 - w_3^*} = \frac{w_3}{1 - w_3} \frac{B_{31} B_{32}}{B_{33} B_{34}}$$

where $B_{3i} = B(a_i^*, b_i^*)$ for $1 \le i \le 4$.

Applying the prior distribution with $w = 0.5, a_1 = b_1 = a_2 = b_2 = 1$, we obtain the posterior probability of H_3 to be $w^* = 5.38(-7)$.

8 Historical Remarks

Much of this work was done under the pressure of responding to a request to provide a document for a legal case. It should not be surprising that a subsequent search of the literature reveals that for this topic which has been studied extensively, much of what has been written above has been presented elsewhere. In particular, papers by G. Barnard (1947) and E.S. Pearson (1947) have anticipated many of the ideas presented here. Barnard introduced the idea that we can select the test statistic so as to minimize the maximum of the P-values over the range (0,1) of p. This led to the CSM test, which he subsequently renounced in favor of the conditionality of the Fisher Exact test. The CSM test was developed in the context of relatively small sample sizes at a time when computers were rather limited. For small sample sizes, the possibility that the maximum is attained for unlikely values of p does not seem very pressing as it is in our case. Pearson was interested in good asymptotic approximations, and incidentally indicated that one should expect our method of using P_F, the P-value of the Fisher Exact test, as the test statistic should yield $P_F(p)$ to be almost independent of p. It is slightly inappropriate to use an abbreviation of Pearson to describe one of the standard tests because of the contributions of Karl Pearson and an abbreviation of Fisher to describe the test statistic pioneered by Karl's son Egon Pearson. Tocher (1950) demonstrated that the test which uses P_F as the test statistic is optimal among *similar* tests of the hypothesis of equality. This implies that the Fisher Exact test is optimal, but does not endorse P_F as the proper P-value to use in interpreting the data.

As computing power grew, various investigators, including Berkson (1978a) criticized the use of the Fisher Exact Test and P_F, suggesting that P_F was too conservative, that the test was not as powerful as it should be, and that a major justification for the use of conditioning on the margins, the ancillarity of the margins, was not justified Berkson (1978b). Yates (1984) responded, defending the application of conditioning on the margins, which he regarded as implicit in the Yates continuity correction as well as the Fisher Exact Test, and claimed that it was obvious to both Fisher and him that there was very little, if any, relevant information in the margins. Yates claimed that the application of Neyman-Pearson theory and power considerations has served to confuse the real underlying issues.

The paper by Yates and the discussion indicates some lack of consensus about the inferential role and definition of the P-value. While the composite nature of our hypothesis H_1 complicates matters a bit and led me to introduce the four measures P_1, P_2, P_3, and P_4, the issue applies even in the case of a *simple* hypothesis. The formal definition of a P-value involves the concept of a test and the use of a test statistic T. Presumably, some decision, terminal or temporary, is to be made depending on whether or not T exceeds some value and the P-value is less than a corresponding value α, in which case the data are regarded as *significant*. That corresponding value is so often taken to be 0.05 or 0.01, for the sake of discussion or to use known tables, that these numbers have assumed undue importance among users of statistical methods. In principle, the appropriate critical P-value should depend in part on the alternative hypotheses and the costs of wrong decisions, and possibly on some background information.

In those cases where the data are not supposed to lead to a terminal decision, the P-value is regarded as a measure of the inferential impact of the data in support of the hypothesis. As such a measure it is incomplete. It says little about the alternative and about the model. In fact the example Fisher presented Barnard, discussed by Barnard in Yates (1984), which led Barnard to accept the conditioning argument and reject CSM, demonstrated exactly this weakness of the inferential interpretation of the P-value. While the conditioning argument is relevant enough so that both E.S. Pearson and Tocher argue in favor of P_F as the appropriate test statistic, one must be careful not to give P_F undue influence as a measure of belief. The fact, that P_F is conditional and technically not the P-value corresponding to a test statistic, is relevant for those who depend on the formal definition of a P-value, but is not ordinarily a major factor in deciding how to interpret it.

Rubin and Stern (1994) present a Bayesian appproach altenative to our measures P_3 and P_4. Assuming H_1 and a prior distribution on the unknown common value of p, they calculate a posterior distrbution using the data (X_1, X_2). This yields a corresponding marginal *predictive* distribution of future values (X_1^*, X_2^*) of the data, in terms of which they calculate $P^* = P(T(X_1^*, X_2^*) \geq T(X_1, X_2)|T(X_1, X_2))$ to serve as a posterior predictive P-value.

At a late date, I was informed of a paper by Berger and Boos (1994) which introduces the idea of using as a P-value the following relative of P_3 and P_4, $P = \sup_{\theta \in \Gamma} p(T;\theta) + (1-\gamma)$ where Γ is a confidence region of confidence γ for a large value of γ, e.g. 0.999.

9 Conclusions

The standard Yates and Pearson methods for testing for the equality of two probabilities do not provide reliable estimates for P-values when one or both sample sizes are moderate, the evidence is strong, and conservative claims are desired. The sources of difficulty are poor asymptotic approximations in the tails of the distributions and the composite nature of the hypothesis to be tested.

One way to deal with this situation is to use exact calculations and to maximize the P-value over a suitable range of values of the nuisance parameter. In this special problem, the nuisance parameter is the supposedly common value of the two probabilities being compared. The appropriate range would be a confidence interval with confidence matched to the P-value attained.

While the Fisher Exact Test yields a conditional P-value, and not a true P-value, using this conditional P-value as a test statistic provides excellent results, since the resulting P-values are almost constant over a wide range of values of the nuisance parameter, and the maximum is relatively small and it is attained at a reasonable value of p.

For more general problems, the use of the likelihood-ratio as a test statistic relieves the statistician of the need to find clever test statistics. This approach works about as well as using the Fisher P-values for the test statistic in our special problem. It applies easily to the problem of testing H_3.

Because the P-value pays insufficient attention to the alternative hypothesis, and the composite nature of the hypothesis to be tested, it is good policy to supplement calculations of the P-value with alternative considerations. Although, the conditional P_F is not a formal P-value, it recommends itself for two reasons. In my case it provides a value which can be presented to a judge as a conservative value. Also it has the advantage of not being influenced by the margins which carry little relevant information.

On the other hand, neither the true P-value nor the conditional value is adequate withut considering, formally or informally, the decision theoretic background of the problem, involving the costs of decisions and prior information from past experience or supplementary data.

Perhaps the result that I found, and probably should not have found, most surprising is the difference between the conventional P-value for the Fisher Exact Test and the one obtained by using it as the test statistic.

Suppose that in my legal problem I had to present one number to reflect the information in the case of the limited data on the first side. In that case I would choose the conditional P-value to present to the judge. Suppose that I had to make a terminal decision in some research effort based on those data. Then I would not depend on the P-value, but would use a decision theoretic approach, based mainly on the likelihood-ratio, and considering costs and some informal prior knowledge. Using a Bayesian approach might be sensible. If I were involved in an ongoing research project where I was required to provide one number as a P-value, I would insist on supplementing one of the P_4 values with the likelihood-ratio, $exp(-Z_L)$, which in this case is 2.51(-3). Finally, if I were not constrained to present a single number, I like $P_F(p)$, supplemented by $p_F^*(Z_F)$, which is 5.11(-1) in our case, where $p_F^*(z) = P(Z_F \geq z|p_1 = \hat{p}_1, p_2 = \hat{p}_2)$. A similar analysis could be done with the likelihood-ratio, yieldng 5.09(-1).

In short, there is no simple solution to the problem of inference, and there is no one number that is adequate to describe the inferential content of the data.

References

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

Berger, R. L. and Boos, D. D. (1994), P-values maximized over a confidence set for the nuisance parameter. Journal of the

American Statistical Association, **89**, 1012-1016.

Berkson, J. (1978a), In dispraise of the exact test. Journal of Statistical Planning and Inference, **2**, 27-42.

Berkson, J. (1978b), Do the marginal totals of the 2×2 tables contain relevant information repecting the table proportions? Journal of Statistical Planning and Inference, **2**, 43-44.

Little, R. J. A. (1989), Testing the equality of two independent Binomial Proportions. The American Statistician, **43**, 283-288.

Pearson, E. S. (1947), The Choice of statistical tests illustrated on the interpretation of tests classed in a 2 × 2 Table. Biometrika, **34**, 139-167.

Rubin, D. B. and Stern, H. S. (1994), Using a posterior predictive check distribution. Latent Variable Analysis, Eds. von Eye and Clogg, London: Sage Publications, 420-438.

Tocher, K. D. (1950), Extensions of the Neyman Pearson theory of tests to discontinuous variates. Biometrika, **37**, 130-144.

Yates, F. (1984), Tests of significance for 2 × 2 contingency tables (with discussion). Journal of the Royal Statisicial Society, series A, **147**, 426-463.