BIOS 6244 Analysis of Categorical Data
October 17, 2005 Lecture

Exact Inference for Small Samples (Sec 2.6)

In our previous consideration of the χ^2, LR, and test for trend, we relied on the χ^2 approximation to calculate p-values for these test statistics. Exact approaches that do not require the χ^2 approximation are also available and are to be preferred, even for moderate to large samples.

Fisher's Exact Test (Sec 2.6.1)

As we saw in our earlier formulation of the problem of testing independence in a 2x2 table, the null hypothesis corresponds to an OR of 1. In 1934, R.A. Fisher proposed a test of this null hypothesis that makes use of the exact distribution of the cell counts (rather than the normal approximation to the binomial, which results in the χ^2 approximation for Pearson's χ^2 test). In order to perform Fisher's exact test, we must consider the *reference set*, consisting of all 2x2 tables with the same row and column totals as the observed table.

Under the Poisson, binomial, or multinomial sampling schemes for the cell counts in a contingency table, the appropriate exact distribution for the cell counts in a 2x2 table is the *hypergeometric*. This distribution is valid *only* if the row and column totals are fixed.

For given row and column marginal totals, the value of n_{11} (or any other cell count in the 2x2 table) determines the other 3 cell counts. Therefore, we need only consider the exact distribution of n_{11}.

In order to be consistent with the original formulation of Fisher's exact test, we must interchange rows and columns in our canonical form for a 2x2 table:

	Exposed	Not Exposed	
Diseased	n_{11}	n_{12}	n_{1+}
Not Diseased	n_{21}	n_{22}	n_{2+}
	n_{+1}	n_{+2}	n

(**NOTE:** The exact p-value for Fisher's exact test is unchanged if the rows and columns of the 2x2 table are interchanged.)

When OR = 1, the probability mass function for n_{11} is given by

$$P(n_{11}) = \frac{\binom{n_{1+}}{n_{11}}\binom{n_{2+}}{n_{+1}-n_{11}}}{\binom{n}{n_{+1}}}, \qquad (5)$$

where the binomial coefficient $\binom{a}{b}$ (read "a choose b") is given by

$$\binom{a}{b} = \frac{a!}{b!(a-b)!}.$$

Recall that in the general case, the hypergeometric distribution is used as a finite population alternative to the binomial:

Let N = population size,

M = # of "successes" in the population,

n' = sample size

W = # of successes in a sample of size n'.

Then

$$\Pr(W = w) = \frac{\binom{M}{w}\binom{N-M}{n'-w}}{\binom{N}{n'}}.$$

In the formulation of Fisher's exact test,

N = n = total sample size,

M = n_{1+} = total # of successes ("diseased" subjects) in the sample,

n' = n_{+1} = total # of subjects in Population 1 ("exposed") in the sample,

W = n_{11} = # of successes ("diseased") among the sample from Population 1 ("exposed").

So, in other words, in Fisher's formulation, we are treating the sample as a small finite population, the "exposed" subjects as a sample from this finite population, and we are inquiring about the probability of obtaining a certain number of "diseased" subjects in our "sample " of exposed subjects.

To test H_0: OR = 1 (i.e., test for independence of rows & columns), we calculate an exact p-value based on the observed value of n_{11}. This is done by summing the hypergeometric probabilities for n_{11} given by Equation (5) over all 2x2 tables in the reference set that are at least as favorable to the alternative hypothesis as the observed 2x2 table is.

Suppose that H_a: OR > 1. Any 2x2 table with the same marginal row and column totals as the observed table that has a count in the (1,1) cell that is greater than or equal to n_{11} in the observed table will be favorable to H_a. The hypergeometric probability for each of these tables should then be included when calculating the upper-tailed p-value.

Fisher's Tea Taster (Sec 2.6.2)

To illustrate his proposed test, Fisher described the following experiment. A colleague of Fisher's claimed that, when drinking tea, she could distinguish whether milk or tea was added to the cup first. To test her claim, Fisher designed an experiment in which his colleague tasted 8 cups of tea; 4 that had milk added first and 4 that had tea added first. The colleague was told that there were 4 cups of tea of each type, and that she should try to select the 4 that had milk added first. The cups of tea were presented to her in random order.

Table 2.8, p. 40, in our text shows the results:

Table 2.8 Fisher's Tea-Tasting Experiment

	Guess Poured First		
Poured First	Milk	Tea	Total
Milk	3	1	4
Tea	1	3	4
Total	4	4	

The null hypothesis H_0: OR = 1 is that Fisher's colleague's guess was independent of the actual order of pouring. The appropriate alternative hypothesis is H_a: OR > 1 since large values of n_{11} indicate that the null hypothesis should be rejected, i.e., that there is a positive association between the true order of pouring and her guess. For this design, the row and column marginal totals *are* fixed since since the colleague knew that 4 cups had milk added first. The distribution of n_{11} under the null hypothesis in this example is the hypergeometric, defined for all 2x2 tables having row and column totals equal to 4. Thus, the possible values of n_{11} are {0, 1, 2, 3, 4}. Table 2.8, in which there are 3 correct guesses of milk added first, has the following probability under H_0:

$$P(3) = \frac{\binom{4}{3}\binom{4}{1}}{\binom{8}{4}} = \frac{\frac{4!}{3!1!}\frac{4!}{1!3!}}{\frac{8!}{4!4!}} = \frac{16}{70} = .229.$$

The only table in the reference set that is favorable to the alternative H_a: OR > 1 that is more extreme than the observed table contains 4 correct guesses. It has $n_{11} = n_{22} = 4$ and $n_{12} = n_{21} = 0$ and probability given by:

$$P(4) = \frac{\binom{4}{4}\binom{4}{0}}{\binom{8}{4}} = \frac{1}{70} = .014 .$$

Table 2.9 contains the entire null distribution of n_{11} for this problem:

Table 2.9 Hypergeometric Distribution for Tables with Margins of Table 2.8

n_{11}	Probability	P-value	X^2
0	.014	1.000	8.0
1	.229	.986	2.0
2	.514	.757	0.0
3	.229	.243	2.0
4	.014	.014	8.0

Note: P-value refers to right-tail probability for one-sided alternative.

Thus, the p-value for testing H_0: OR = 1 vs. H_a: OR > 1 for Fisher's colleague is P(3) + P(4) = .014 + .229 = .243, i.e., we accumulate the hypergeometric probabilities for all 2x2 tables with row and column margins equal to 4 that are favorable to H_a.

This p-value provides no evidence against the null hypothesis of independence. Therefore, Fisher's colleague does not appear to be able to correctly determine whether milk or tea was poured first. However, note that the small sample size severely limits the ability of Fisher's exact test to detect a departure from the null hypothesis. The only outcome leading to rejection of H_0 would have been $n_{11} = 4$ (p = .014).

P-Values & Type I Error Probabilities (Sec 2.6.3)

Recall that in applying the χ^2 and LR tests of independence, the alternative hypothesis was H_a: OR $\neq$ 1.

For a 2-tailed alternative, we calculate the p-value for Fisher's exact test by accumulating the hypergeometric probabilities of all tables no more likely than the observed table; that is, one adds the probabilities of all possible outcomes y such that $P(y) \leq P(n_{11})$, where n_{11} is the observed count. Using the probabilities in Table 2.9 for the tea-tasting example, we sum all probabilities that are no greater than P(3) = .229. This yields

2-tailed p-value = P(0) + P(1) + P(3) + P(4) = .486.

If the marginal totals for either the rows or columns are equal, then the hypergeometric distribution is symmetric. In this case, there is a shortcut for calculating the 2-tailed p-value: we simply calculate the appropriate 1-tailed p-value and then double it. So the 2-tailed p-value for Fisher's exact test in this case is

$$\text{p} - \text{value} = \left\{ \begin{array}{l} 2 \text{ x upper} - \text{tailed p} - \text{value if } n_{11} \geq n_{12} \\ 2 \text{ x lower} - \text{tailed p} - \text{value if } n_{11} \leq n_{12} \end{array} \right\}$$

In the tea-tasting example, $n_{11} = 3 > 1 = n_{12}$, so the 2-tailed p-value = 2 x .243 = .486.

What if we applied the χ^2 test to the data in Table 2.8? We see that in Table 2.9, Agresti has calculated the χ^2 test statistic corresponding to each of the possible values of n_{11}. The exact 2-tailed p-value for the χ^2 test is calculated in the same way as for Fisher's exact test; i.e., we accumulate the probabilities of all tables for which the χ^2 test statistic is at least as large as the observed value ($X^2 = 2$). These probabilities are calculated using the hypergeometric probabilities given in Table 2.9:

x^2	$\Pr(X = x^2)$
0	.514
2	.229 + .229 = .458
8	.014 + .014 = .028

Figure 2.2, p. 42, contains a graph of the exact distribution of X^2 for the data in Table 2.8:

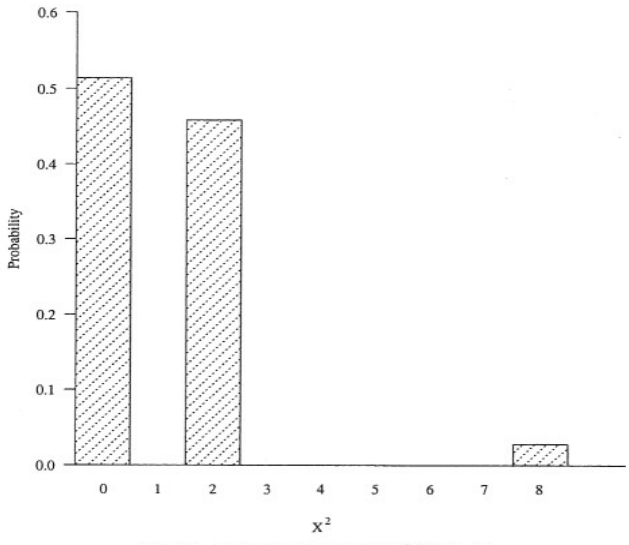

Figure 2.2 Exact distribution of Pearson X^2 for Table 2.8.

For the data in Table 2.8, $X^2 = 2.0$. So the exact upper-tailed p-value for the χ^2 test is .458 + .028 = .486. Note that this is the same as the exact 2-tailed p-value for Fisher's exact test. *For 2x2 tables, the exact p-values for the χ^2, LR, and Fisher's exact test will be identical.*

As the row and column totals in the contingency table increase, the calculations for the hypergeometric probabilities associated with Fisher's exact test become more and more tedious since the reference set consists of *all* possible tables with same row and column totals as the observed table. Methods for approximating the p-value for Fisher's exact test have been proposed, but these are no longer necessary since modern statistical packages can handle the exact p-value calculations.

(We will discuss how to calculate exact p-values for Fisher's exact test in the Computer Lab session on October 19.)

For small sample sizes, the exact distribution calculated using Equation (5) is highly discrete, in the sense that the number of possible values for n_{11} is relatively small. As a result, the exact p-value also has only a relatively small number of possible values. For example, for the data in Table 2.8, there are only 5 possible p-values for a 1-sided Fisher's exact test and only 3 possible p-values using the exact version of the χ^2 test.

This discretenessof p-values has an impact on the Type I error rate (i.e., the significance level, α) for the exact test. Suppose we want to use $\alpha = .05$ when testing the null hypothesis, i.e., reject H_0 if $p < .05$. However, because of the discreteness of the test statistic, it is usually not possible to conduct an exact test that has α exactly equal to .05. Recall that $\alpha = \Pr(\text{reject } H_0 \mid H_0 \text{ is true})$. To perform Fisher's exact test, we want to choose a rejection region R so that $\Pr(n_{11} \in R \mid OR = 1) = .05$. However, from Table 2.9, we see that, for a 2-tailed test, the rejection region that yields α closest to (but less than) .05 is {0, 4}. That is, $\Pr(n_{11} = 0 \text{ or } 4) = P(0) + P(4) = 2(.014) = .028$. All other choices for R yield a value of $\alpha > .05$. For a 1-tailed test, the closest we can get is $\alpha = .014$.

Tests for which the true value of α is less than the nominal (or desired) value of α (typically .05 in clinical research) are called *conservative*. Tests for which the true value of α is greater than the nominal value are called *liberal*.

The conservativeness of exact tests for discrete test statistics can be further illustrated by thinking of the p-value itself as a random variable. For test statistics having a continuous distribution (like the χ^2), it can be shown that the p-value has a uniform distribution over the interval [0,1] if the null hypothesis is true. In other words, the p-value is equally likely to fall anywhere between 0 and 1 and the probability that the p-value falls between any two values *a* and *b* = *b*-*a*. So, for example, the probability that the p-value falls below .05 is equal to .05 – 0 = .05. Furthermore, the expected p-value under the uniform distribution is .5. For test statistics having a discrete distribution, the p-value also has a discrete distribution and, under the null hypothesis, it typically has an expected value greater than .5. For example, in the tea-tasting example, we can calculate the expected 1-tailed p-value using the values in Table 2.9:

$$E(\text{p-value}) = 1(.014) + .986(.229) + .757(.514) + .243\ (.229) + .014\ (.014) = .685 > .5.$$

So, the exact p-value for Fisher's exact test for the tea-tasting experiment tends to be too large "on average."

To help diminish the effect of the conservativeness of exact tests for discrete data, one can use a slightly different version of the p-value, called the *mid p-value*. This is equal to the appropriate exact p-value, minus half the point probability of the observed value of the test statistic. It has an expected value of .5 under the null hypothesis, as do p-values for any test statistic having a continuous distribution. For the tea-tasting experiment, using the values in Table 2.9, we find the 1-tailed mid p-value for Fisher's exact test to be $.243 - \frac{1}{2}(.229) = .129$, compared with .243 for the exact p-value. For the 2-tailed exact test based on the χ^2 test statistic, the mid p-value is $.486 - \frac{1}{2}(.458) = .257$, compared with .486 for the exact p-value. (Note that the 2-tailed mid p-value can be obtained by doubling the 1-tailed mid p-value.) For the tea-tasting data, the exact p-value and mid p-value both yield the same conclusion, but this will not always be the case.

Unlike the exact p-value calculated for a discrete test statistic, using the mid p-value does not guarantee that the true significance level will be less than or equal to the nominal level α. (See Problem 2.27, p. 50). However, the mid p-value-based test usually performs well, and is less conservative than the test based on the exact p-value.

In Fisher's tea-tasting experiment, both the row and column marginal totals were fixed. In Epi studies, the marginal totals for rows and/or columns may be random. Fisher's exact test can still be applied in these situations by conditioning on the observed row and column totals, thereby treating them as fixed. This "conditional" version of Fisher's exact test has been shown to also perform well. "Unconditional" test procedures that take the randomness of the row and/or column into account are available in sophisticated statistical packages such as StatXact, but are not considered to have any particular advantage over Fisher's exact test when the mid p-value is used.

<u>Exact Confidence Interval for the Odds Ratio (Sec 2.6.4)</u>

In Section 1.3.3, we discussed finding an exact CI for a binomial proportion π. The approach that was recommended there was to include in the 95% CI(π) all values π_0 such that H_0: $\pi = \pi_0$ would not be rejected using $\alpha = .05$. We recommend the same approach in finding an exact 95% CI(OR). The issue of "conservativeness" also arises here since an exact 95% CI(OR) will tend to have a true confidence coefficient larger than 95%, i.e., our intention is to construct an exact 95% CI(OR), but we may actually wind up with a 98% CI(OR). Basing the CI on those values π_0 for which H_0: $\pi = \pi_0$ would not be rejected using the mid p-value approach will reduce this conservatism, although the true confidence coefficient may actually be less than 95%.

Example (tea-tasting, cont.)

exact 95% CI(OR): (.21, 626.17)

mid p-based 95% CI(OR): (.31, 308.55)

Note the extreme width of both CI's, owing to the small sample size (n = 8).

The exact CI(OR) is available in SAS and SPSS and the mid p-based CI is available in StatXact.

Exact Tests of Independence for I x J Tables (Sec 2.6.5)

A generalization of Fisher's exact test, called the *Fisher-Freeman-Halton test*, is available for I x J tables with more than 2 rows or columns. The *multivariate hypergeometric* distribution is used to calculate p-values for this test. There are also exact versions of the χ^2 and LR tests available for these larger tables.

NOTE: The exact p-values for the χ^2, LR, and Fisher's exact test do not agree for general I x J tables, only for 2x2.

Example (Oral Lesions in India)

The following data on the location of oral lesions were obtained in house-to-house surveys in 3 geographical regions of rural India. The question of interest is whether the distribution of the site of oral lesions differs significantly among the 3 regions. Neither the rows nor columns of this table can be ordered in any meaningful way.

Oral Lesions Data Set

Site of Lesion	Kerala	Gujarat	Andhra
Labial Mucosa	0	1	0
Buccal Mucosa	8	1	8
Commissure	0	1	0
Gingiva	0	1	0
Hard Palate	0	1	0
Soft Palate	0	1	0
Tongue	0	1	0
Floor of Mouth	1	0	1
Alveolar Ridge	1	0	1

The following results were obtained:

Test	χ^2 Approximate p-value	Exact p-value	Mid p-value
χ^2	.140	.027	.027
LR	.106	.036	.035
Fisher-Freeman-Halton	--	.010	.008

Note that the mid p adjustment has little effect on the exact p-values.

Recall that if the row variable and/or column variable of an I x J table is ordinal, the methods described in Sec. 2.5 are preferable to the Fisher-Freeman-Halton test.