CONTINGENCY TABLES ARE NOT ALL THE SAME

David C. Howell
University of Vermont

To most people studying statistics a contingency table is a contingency table. We tend to forget, if we ever knew, that contingency tables can be formed in different ways, and how the table is restricted may influence the analysis we run.

The Mathematics of a Lady Tasting Tea

Let's start with a very famous example from R. A. Fisher. This is often known as the "Lady Tasting Tea" example, and, according to Fisher's daughter, it is a true story. The basic idea is that one day when people who worked around Fisher were having afternoon tea, one of them, Muriel Bristol, claimed that she could tell whether the milk was added to the cup before or after the tea. Fisher immediately turned this into an experiment by preparing eight cups of tea, four of which had milk added first and four of which had milk added second. He then put the cups in front of Muriel and asked her to identify the four cups that had milk added first. By the way, Muriel was no slouch. She was a Ph.D. scientist, back in the days when women were not Ph.D. scientists and she established the Rothamstead Experiment Station in 1919. This was the place that Fisher was later to make famous. I think you should know what Muriel looked like, so here she is—stolen without permission from Wikipedia. (No, that isn't Einstein, although it looks like him.)

This was a great example for Fisher, and we'll come back to it later, but first I want to modify the experiment to include 96 cups instead of 8. Assume that each day for 12 days Muriel was presented with eight cups, four of each kind, and asked to make the identification. (Also assume that observations across days were independent, which seems like a reasonable assumption.) The data are then collapsed over all 12 days. (I made this change to have a much larger total sample size.)

Without looking at the data you know something about them. There will be 48 cups with milk first and 48 cups with milk second. In addition, because of the instructions to Muriel, there will be 48 guesses of First and 48 guesses of Second. Finally, there will be 96 total observations.

		True Condition		
		First	Second	
Guess	First	29	19	48
	Second	19	29	48
		48	48	96

I created these data with these cell frequencies so as to have a table whose Pearson chi-square statistic will be significant at close to α = .05. In this case the probability under the null is .0412. (These data are more extreme that the actual data as far as Muriel's ability to detect differences is concerned.)

Pearson's Chi-Square Test

Let's start with a simple Pearson chi-square test for a 2 × 2 table, along with the odds ratio.

$$\chi^2 = \Sigma\frac{(O-E)^2}{E} = \frac{(29-24)^2}{24} +$$

$$\ldots + \frac{(29-24)^2}{24} = 4.17$$

$$OR = \frac{29/19}{19/29} = \frac{1.5263}{01.6552} = 2.33$$

As I said, this value of chi-square is significant at $p = .0412$. We can reject the null hypothesis and conclude that Muriel's response and the way the tea was prepared are not independent. In other words she guessed correctly at greater than chance levels. (Note that I did not include Yates' correction for these data, but if I had the chi-square would have been 3.38 with a p value of .0662. We will come back to that later.)

Fisher's Exact Test and the Hypergeometric Distribution

Fisher specifically did not evaluate these data with a Pearson chi-square, and not just because he couldn't stand Pearson, which he couldn't. His reasoning involved the hypergeometric distribution. When both row and column totals (the "marginals") are fixed, the hypergeometric distribution will tell us the probability that we will have a specified number of observations in cell$_{11}$. (We could have picked on any cell, but cell$_{11}$

seems like a nice choice. We only need to worry about one cell because we only have 1 *df*. If we had a 10 in cell$_{11}$, then cell$_{12}$ must be 38, cell$_{21}$ but be 28, and cell$_{22}$ must be 10.)

The formula for the hypergeometric, if you must know, is

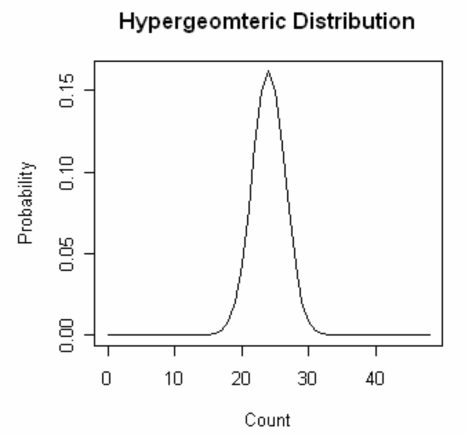
$$p(x) = \frac{\binom{n_{1.}}{x}\binom{n_{2.}}{n_{.1}-x}}{\binom{N}{n_{.1}}} = \frac{\left(\binom{48}{29}\binom{48}{19}\right)}{\binom{96}{48}} = \frac{\left(\frac{48!}{29!19!}\right)\left(\frac{48!}{19!29!}\right)}{\frac{96!}{48!48!}} = .020701$$

Where x = observation in cell$_{11}$, $n_{1.}$ = total of row 1, $n._{.2}$ = total of column 2, and N = total number of observations. But of course this will only give us the probability of exactly 29 observations in cell$_{11}$. We are going to want the probability of 29 *or more* observations in that cell, so we have to evaluate this expression for all values of 29 and higher. This is shown below, though I only carried 5 decimals and the entries rapidly drop to 0.00. But to make this a two-tailed test we also need to know the probability of 19 *or fewer* observations, which, because we have the same number of each type of tea, is the same. So the two-tailed probability is .0656.

Number in Upper Left	Probability
29	.02070
30	.00830
31	.00280
32	.00079
33	.00018
34	.00005
39	.00000
40	.00000
…	…
48	…
Sum	**.03282**

Number in Upper Left	Probability
19	.02070
18	.00830
17	.00280
16	.00079
15	.00018
14	.00005
13	.00000
12	.00000
…	…
0	.00000
Sum	**.03282**

In case you are interested, the distribution is plotted below, where you can see that the vast majority of outcomes lie between about 18 and 31.



Using Fisher's Exact Test we have a one-sided probability of .033, which would lead us to reject the null hypothesis. The two-sided probability would be .0656, which would not allow us to reject the null. Notice that we did not reject the (two-tailed) null hypothesis using Fisher's test but we did using the Pearson chi-square test. This is most likely a result of the fact that observations are discrete, whereas the chi-square distribution is continuous.

I earlier gave the "corrected" version of chi-square, called Yates' correction, which had a probability of .0662. That number is very close to Fisher's value, which is as it should be. Yates was trying to correct for the continuousness of the chi-square distribution, and he did so admirably.

Contingency Tables with One Set of Fixed Marginals

In the data table that we just examined, both the row and marginal totals were fixed. We knew in advance that there would be 48 cups of each type of tea and that Muriel would make 48 choices of each type. But consider a different example where the row totals are fixed but not the column totals. I will take as an example Exercise 6.13 from the 6th edition of *Statistical Methods for Psychology.* In 2000 the State of Vermont, to their very great credit,—a slight editorial comment—approved a bill authorizing civil unions between gay and lesbian partners. The data shown below suggest that there was a difference on this issue between male and female legislators. I chose this example, in part, because the sample size (145 legislators) is reasonable and the probability (p = .019) is significant but not extreme. Notice that the row totals are fixed because if someone demanded a recount the number of male and female legislators would be the same (44 and 101) but the number of votes for and against the legislation could change. The calculations of Pearson's chi-square and the odds ratio are straightforward.

	Vote		
	Yes	**No**	**Total**
Women	35	9	44
Men	60	41	101
Total	95	50	145

$$\chi^2 = \Sigma\frac{(O-E)^2}{E} = \frac{(35-28.83)^2}{28.83} + \ldots + \frac{(41-34.83)^2}{34.83} = 5.502$$

$$OR = \frac{35/9}{60/41} = \frac{3.889}{1.463} = 2.66$$

The usual way to evaluate this test statistic is to compare it to the mathematically defined chi-square distribution. Although that distribution will not be an exact fit to the sampling distribution of this statistic, it will be very close. Using R, or any other statistical software, the two-tailed *p* value is .01899.

By normal standards, this value is statistically significant and we can conclude that how legislators voted depended, to some extent, on their gender. Women were more likely than men to vote for the legislation. The odds ratio was 2.66. Notice that this is a two-tailed test because we could have a large value of chi-square if either men outvoted women or women outvoted men—the difference is squared.

Another, and equivalent, way of running this test is to notice that 82% of the female legislators supported the measure whereas on 59% of the male legislators did. You probably know from elsewhere that we can test the difference between two proportions using *z*

$$p = (35+60)/145 = .6552$$

$$z = \frac{p_1 - p_2}{\sqrt{p(1-p)(\frac{1}{n_1}+\frac{1}{n_2})}} = \frac{.7954-.5941}{\sqrt{.6552(.3448)(.0326)}} - \frac{.2013}{.0858} = 2.346$$

The two-tailed probability is .0190, which is what we found with chi-square. (The reason that they agree so well is that when you have 1 *df* a chi-square distribution is just the square of a normal distribution, so this is really the same test.)

Fisher's Exact Test

Although the marginal totals on the columns in this example are not fixed, we could act as if they are and apply Fisher's Exact Test. If we did so we would find a probability under the null of .0226, which is slightly higher than we found with chi-square.

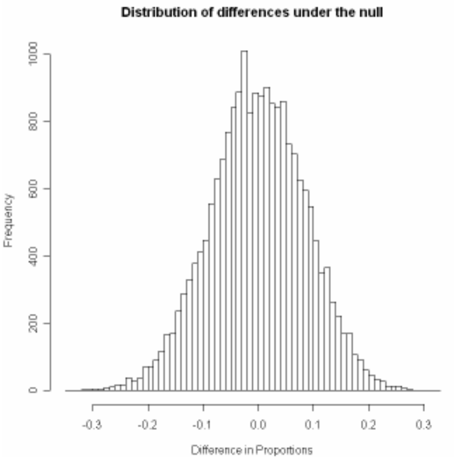
Resampling

But let's look at this still a third way. Even though we have a lot of cases (145) the distribution is still discrete (χ^2 can only take on a specific set of different values) and is being fit against a continuous distribution. So suppose that we set up a simple resampling study for the case with one fixed set of marginals.

First we will assume that the null hypothesis is true so we are sampling from populations with the same proportion of Yes votes. Then our best estimate of the common population proportion is 95/145 = .6552. In our resampling study we will draw 44 cases, corresponding to women in the legislature. For each woman we will draw a random number between 0 and 1. If that number is less than .6552 we will record her vote as Yes. After we have drawn for all 44 women we will compute the proportion of them who voted Yes. (I could have accomplished the same thing by drawing 44 observations from a binomial with p = .6552—e.g. `rbinom(n = 1, size = 44, prob = .6552)`. I did things the long way because it makes it easier to grasp the underlying process.) In the long run we would expect to have 65.52% of women voting Yes, if the null hypothesis is true, but the actual counts will vary due to normal sampling error. Then we will do the same thing for males, but this time making 101 draws with number of Yes votes equal to the number of times our random number was greater than .6552. We will record the difference between male and female Yes votes. We will then repeat this "experiment" 10,000 times, each time generating the number of votes for the legislation by both men and women. Notice that we have set the common probability in both cases at .6552, so, on average, the differences between males and females will come out to 0.00. When we are all done we will have the distribution of differences for 10,000 cases based on a true null hypothesis, and we can ask what percentage of those differences exceeded .2013, the difference in proportions that we found from our study[1]. Notice that I am holding the number of males and females constant, but allowing the votes to vary.

A histogram of the results is printed below.

[1] Although I am using the difference between the two sample proportions as my statistic, I could instead calculate chi-square for each resampling and use that as my statistic. I just need something that represents a measure of how similar or different are the behaviors of men and women. (In the program that is attached, I use chi-square as my statistic, but I do not compare it to the chi-square distribution.)



The proportion of differences greater than .2013 is 0.0079
The proportion of differences less than -.2013 is 0.0105

If we add together the two tails of the distribution we find that 1.84% of the observations exceeded ± .2014, which was our obtained difference. These results are comforting because they very nearly duplicate the results of our chi-square test, which had a *p* value of .190. One important thing that this tells us is that chi-square is an appropriate statistic when one set of marginals are fixed—at least if we have relatively large sample sizes**.**

The Case of No Fixed Marginals

Now we will carry the discussion of contingency tables one step further by considering results in which we would not be willing to assume that either set of marginals is fixed. There have been a number of studies over the years looking at whether the imposition of a death sentence is affected by the race of the defendant (and/or the race of the victim). Peterson (2001) reports data on a study by Unah and Borger (2001) examining the death penalty in North Carolina in 1993-1997. The data in the following table show the outcome of sentencing for white and nonwhite (mostly black and Hispanic) defendants when the victim was white. The expected frequencies are shown in parentheses.

Sentencing as a function of the race of the defendant

	Death Sentence		
Defendant's Race	**Yes**	**No**	**Total**
Nonwhite	33 (22.72)	251 (261.28)	284
White	33 (43.28)	508 (497.72)	541
Total	66	759	825

This table is different from the others we have seen because if we went out and collected new data for some other time period or some other state, both the row totals and the column totals would be expected to change. Although it would be possible to do the arithmetic for Fisher's Exact Test, it would be hard to argue that we should condition on these marginals. So let's look at a different way of approaching the problem.

First of all there is nothing to stop us from running a plain old Pearson chi-square, and, in fact, that is probably what we should do. These calculations follow.

$$\chi^2 = \Sigma \frac{(O-E)^2}{E}$$

$$= \frac{(33-22.72)^2}{22.72} + \frac{(251-261.28)^2}{261.28} + \frac{(33-43.28)^2}{43.28} + \frac{(508-497.82)^2}{497.72}$$

$$= 7.71$$

The probability of chi-square = 7.71 on 1 *df*, if the null is true, is .0055, leading us to reject the null. (The probability for Fisher's test, which I think is not appropriate here, is .007.)

The Resampling Approach Again

Fisher met a lot of resistance to his idea on contingency tables because people objected to holding the marginal totals fixed. They argued that if the experiment were held again, we would be unlikely to have the same numbers of white and black defendants. Nor would we have the same number of death sentences. (This argument actually went on for a very long time, and we still don't have good resolution. I, myself, have moved back and forth, but I now think that I would recommend Fisher's Exact Test only for the case of fixed marginals.)

One of the huge problems in this debate was that the calculations became totally unwieldy if you let go of the fixed marginals requirement. BUT now we have computers, and they like to do things for us. Computers, like Google, are our friends! as you saw in the previous problem. Suppose that we pose the following task. We have 825 defendants. For this sample, 284 of them are nonwhite and 541 of them are white. Thus 284/825 = .2442 is the proportion of nonwhite defendants. Similarly, 66/825 = .08 is the proportion of defendants who were found guilty. From now on I don't care about how many nonwhite or guilty defendants we will have in any sampling, but if the null hypothesis is true, .2442 * .08 = .0195 is the expected proportion of nonwhite guilty defendants we would have in any sample. So what we can do is draw a sample of 825 observations where the probability of landing in cell_{11} is .0195. Similar calculations will give us the expected proportions landing in the other cells. Notice that these proportions are calculated by multiplying together the proportions of each race and the proportions of each verdict. This would only be the case if the null hypothesis were true.

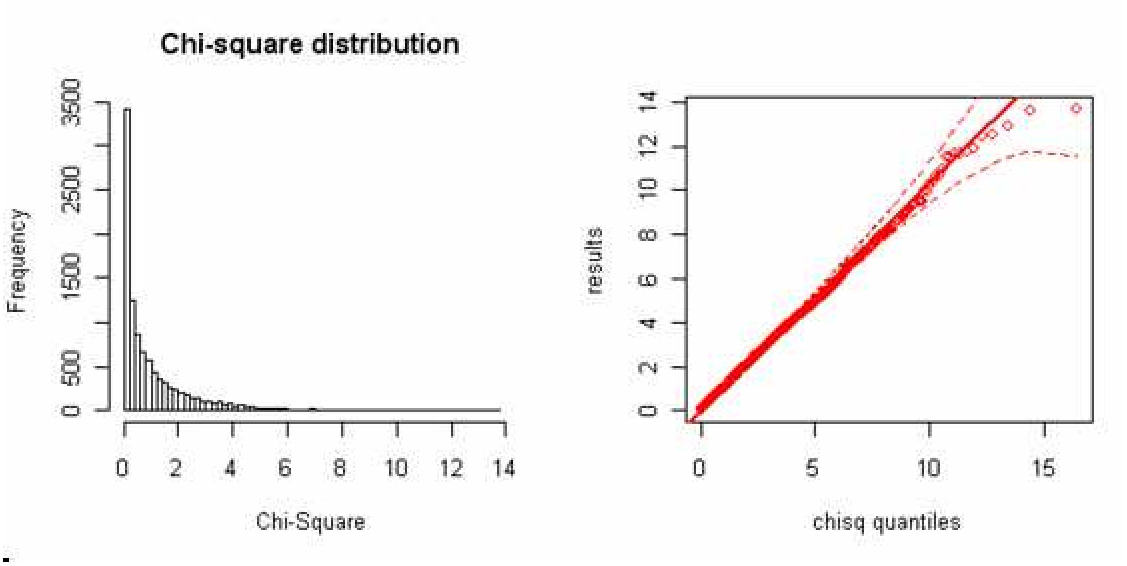
Instead of looking at all possible outcomes and their probability, which would be a huge number, I'll instead take a random sample of all of those outcomes. Let's run 10,000 experiments. Each time we run an experiment we let a random number generator put the observations in the four cells randomly based on the calculated probabilities under the null. For example, the random number generator (drawing from the set of numbers 1-4) might pick a 1 with a probability of .0195 and thus add the defendant to the 1st cell, which is $\text{cell}_{1,1}$, which is the cell for NonWhite/Yes. The program makes 825 of these random assignments and pauses. We have filled up our contingency table with one possible outcome of 825 observations, but we need some way to tell how extreme this particular outcome is. One way to do this is to calculate a chi-square statistic on the data. That is easy to do, and we calculate this statistic for each particular sample. *Keep in mind that we are using chi-square just as a measure of extremeness, not because we will evaluate the statistic against the chi-square distribution. We could have used the difference between the cross products in the matrix.* The more extreme the results, the larger our chi-square. We just tuck that chi-square value away in some array and repeat the experiment all over again. And we do this 10,000 times.

I actually carried out this experiment. (That took me 37 seconds, which is faster than I could calculate Pearson's chi-square once by hand.) For my 10,000 resamples, the proportion of cases that were more extreme than a chi-square of 7.71 was .0056, which is in excellent agreement with the chi-square distribution itself, which gave a probability of .0055. Remember that these data were drawn at random, so the null hypothesis is actually true. This is beginning to look as if the chi-square distribution is an excellent approximation of the results we obtained here, even when neither set of marginals is fixed. We can look more closely at the results of the resampling study to see if this is really true.

The results of this sampling procedure are shown below. What I have plotted is not the mathematical chi-square distribution but the distribution of chi-square values from our 10,000 resampling. It certainly looks like a chi-square distribution, but looks can be deceiving.

One nice feature of the results is that the mean of this distribution is 0.98 with a variance of 1.89. (On a second run those values were 1.022 and 2.0027, respectively.) For the true chi-square distribution the mean and variance are *df* and 2*df*, or 1 and 2, which is very close. Let's see how well they agree across a range of values.

In *Statistical Methods for Psychology, 7th ed.* I used QQ plots to test normality. I just plotted the obtained quantiles against the ones expected from a normal distribution. I will do the same thing here except that I will plot my obtained chi-square values against the quantiles of the chi-square distribution. For example, I would expect 1% of a chi-square distribution with 1 *df* to exceed 6.63. Actually I found 94 / 10,000 = .94% at 6.63 or above. I would expect 50% of the results to be greater than or equal to 0.45, and actually 49.8% met that test. If I I obtain similar pairs of values for all of the percentages between 0 and 100, I would have the following result.



Notice that the values fall on a straight line. There are only trivial deviations of the results from that line. (The dashed lines represents confidence intervals.)

So now we have at least four ways to evaluate data in 2 × 2 contingency tables. We can run a standard Pearson's chi-square test (or a likelihood ratio test, which I have not described), we can assume that both sets of marginals are fixed and use Fisher's Exact test, we can assume that row marginals are fixed but not the column marginals, or we can assume nothing about the marginals (other than the total number of observations) and run a resampling test. In the language of statisticians we are moving from the hypergeometric to the binomial to the multinomial distributions. The fact our results generally come out to be close is encouraging. I'm a great fan of randomization tests, but that is partly because I like to write computer programs. For those who don't like to write programs, you can use either Fisher's Exact Test or Pearson's chi-square. There is a slight bias toward Fisher's test in the literature when you have small sample sizes, but don't settle for it until you have read the rest of this document.

The one thing that I would not recommend is using Yates' Correction with the standard Pearson chi-square. If you are worried about discreteness of the probability distribution go with Fisher's Exact Test.

Remember that Fisher's Exact Test applies only to 2 × 2 tables. It can be expanded to larger tables, (Howell and Gordon, 1976), but that is not commonly done. (In R and S-Plus there are no restrictions on dimensionality if you use "fisher.test( )

But are Things Really That Good??

It's nice to know that with reasonably large samples the randomization tests (and the hypergeometric) produce results very similar to those produced by chi-square. But what if we don't have large samples? In that case the chi-square distribution may not be an adequate fit to the resulting test statistic. I will start with Fisher's original experiment, which only had 8 cups of tea. His results are below.

		True Condition		
		First	Second	
Guess	First	3	1	4
	Second	1	3	4
		4	4	8

Suppose that we compute chi-square on these data. Our result will yield of chi-square of 2.00 on 1 *df* with an associated probability of .1573. But Fisher's exact test gives a probability of .4857!!! The reason is fairly simple. There are only a few ways the data could have come out. The table could have looked like any of the following.

	4 0	3 1	2 2	1 3	0 4
	0 4	1 3	2 2	3 1	4 0
χ^2 =	8.00	2.00	0.00	2.00	8.00
p	.005	.157	1.00	.157	.005
Fisher 2-tail	.029	.486	1.00	.486	.029
Fisher 1-tail	.014	.243	.757	.986	1.00

Notice how discrete the results are and how far the (correct) Fisher probabilities are from the chi-square probabilities. With small samples and fixed marginals I strongly suggest that you side with Fisher—he could use some company. We know that for fixed marginals his values are theoretically correct, and we now know that chi-square probabilities are not even close.

So the *p* values assigned by Fisher and by the chi-square distribution are greatly different. But perhaps this is a bad example because the frequencies are so very small. So let's go back to the example of the Vermont legislature and cut the sample sizes in each cell by approximately 5. (I had to cheat a bit to get whole numbers.) Remember that this is not a case where we would prefer to use Fisher's Exact test because only the row totals are constant. So I will compare a standard chi-square test with the resampling procedure that we used earlier.

	Vote		
	Yes	**No**	**Total**
Women	7	2	9
Men	12	8	20
Total	19	10	29

$$\chi^2 = \Sigma\frac{(O-E)^2}{E} = \frac{(7-3.724)^2}{3.724} + \ldots + \frac{(8-6.896)^2}{6.896} = 0.868$$

$$OR = \frac{7/2}{12/8} = \frac{3.5}{1.5} = 2.33$$

The *p* value (two-sided) from chi-square = .351, which is not significant. If we had asked for Yates correction, chi-squared would have been .2597 with a *p* value of .61 whereas Fisher's test would have had a *p* value of .43. Notice that these values are all over the place.

Now let's use the same resampling approach that we used earlier, holding the row marginals fixed but not the columns. The obtained difference in the two proportions were .7778 - .6000 = .1778. The probability of an outcome more extreme than this, in either direction, was .1703 + .1818 = .3521, which is virtually the same as the chi-square probability, but quite different from the probability given by Fisher's test (.43). I prefer the resampling approach because it is in certain ways an exact test. But notice that the standard chi-square test produces almost the same statistic. This tells me that chi-square is a good fallback when you don't want to do a randomization test.

Now let's go back to the death sentence data where neither set of marginals are fixed. If I cut my cell frequencies by approximately 9, I obtain the following data. The total sample sizes are larger than I would like for an example, but I can't reduce the cells by much more and still have useful data.

	Death Sentence		
Defendant's Race	**Yes**	**No**	**Total**
Nonwhite	4	28	32
White	4	56	60
Total	8	84	92

The standard Pearson chi-square on these data is 0.8944 for p = .3443. When we run a randomization test with the null hypothesis true and no constraints on the marginal totals we find 3518 results that are greater than our obtained chi-square, for p = .3518. Notice how well that agrees with the Pearson chi-square. Yates correction gave a p value of .5773, whereas Fisher's Exact Test gave a p = .4422. Again we see that the standard chi-square test is to be preferred. (Well, I later reduced my cells by a factor of approximately 10 and found a chi-square probability of .395 and a resampling probability of .410—who can complain about that?)

So what have we learned?

Well, I have learned a lot more than I expected. When the data can reasonably be expected to have both sets of marginal totals fixed, then conditioning on those marginals, which is what Fisher's Exact Test does, is the preferred way to go. However when one or both sets of marginals are *not* fixed, and when the sample size is small, Fisher's test gives misleading values. Both when one set of marginals is fixed and when no marginals are fixed, the standard Pearson chi-square test, without Yates correction, is to be preferred. This would appear to be in line with the recommendation given by Agresti (2002), which is always reassuring. The chi-square test (when one or zero marginals are fixed) agrees remarkably well with randomization tests that seem to be reasonable ways of conceiving of the data. I had expected to find something like this, but I never thought that it would be anywhere as neat as it is. Now I have to go back to the book that I am revising and re-revise the section on Fisher's Exact Test.

The Programs

The programs that I created with R can be easily downloaded. They are not particularly elegant, but they work just fine. In each of them I have several sets of data, with a "#" in front of all but one. A # simply comments out the line. You can create your own data matrix by mimicking what I have there.

Plot the Hypergeometric Distribution
www.uvm.edu/~dhowell/StatPages/More_Stuff/Chi-square/hypergeometric.r

---

Row Marginals Fixed
www.uvm.edu/~dhowell/StatPages/More_Stuff/Chi-square/**rowfixed.r**

No Fixed Marginals

www.uvm.edu/~dhowell/StatPages/More_Stuff/Chi-square/NoMarginalsFixed.r

##

8/15/2009