Common Univariate and Bivariate Applications of the Chi-square Distribution©

The probability density function defining the chi-square distribution is given in the chapter on Chi-square in Howell's text. Do not fear, we shall not have to deal directly with that formula. You should know, however, that given that function, the mean of the chi-square distribution is equal to its degrees of freedom and the variance is twice the mean.

The chi-square distribution is closely related to the normal distribution. Image that you have a normal population. Sample one score from the normal population and compute $Z^2 = \frac{(Y-\mu)^2}{\sigma^2}$. Record that Z^2 and then sample another score, compute and record another Z^2, repeating this process an uncountably large number of times. The resulting distribution is a chi-square distribution on one degree of freedom.

Now, sample two scores from that normal distribution. Convert each into Z^2 and then sum the two scores. Record the resulting sum. Repeat this process an uncountably large number of times and you have constructed the chi-square distribution on two degrees of freedom. If you used three scores in each sample, you would have chi-square on three degrees of freedom. In other words, $\chi_n^2 = \sum_{i=1}^{n} Z^2 = \sum \frac{(Y-\mu)^2}{\sigma^2}$.

Now, from the definition of variance, you know that the numerator of this last expression, $\sum (Y-\mu)^2$, is the sum of squares, the numerator of the ratio we call a variance, sum of squares divided by *n*. From sample data we estimate the population variance with sample sum of squares divided by degrees of freedom, (*n* - 1). That is, $s^2 = \frac{\sum (Y-\bar{Y})^2}{n-1}$. Multiplying both sides of this expression by (*n* - 1), we see that $\sum (Y-\bar{Y})^2 = (n-1)s^2$. Taking our chi-square formula and substituting $(n-1)S^2$ for $\sum (Y-\mu)^2$, we obtain $\chi^2 = \frac{(n-1)s^2}{\sigma^2}$, which can be useful for testing null hypotheses about variances. You could create a chi-square distribution using this modified formula -- for chi-square on (*n* - 1) degrees of freedom, sample *n* scores from a normal distribution, compute the sum of squares of that sample, divide by the known population variance, and record the result. Repeat this process an uncountably large number of times.

Given that the chi-square distribution is a sum of squared *z*-scores, and knowing what you know about the standard normal distribution (mean and median are zero), for chi-square on one *df*, what is the most probable value of chisquare (0)? What is the smallest possible value (0)? Is the distribution skewed? In what direction (positive)?

Now, consider chi-square on 10 degrees of freedom. The only way you could get a chi-square of zero is if each of the 10 squared *z*-scores were exactly zero. While zero is still the most likely value for *z* from the standard normal distribution, it is not likely that you would get all 10 scores exactly equal to zero, so zero is no longer the most likely value of chi-square. Also, given that 10 squared *z*-scores go into that chi-square its mean increases (from 1 to 10). Note also that the positive skewness of the chi-square distribution decreases as the *df* increase.

It is strongly recommended that you complete the SAS lesson, "Simulating the Chi-Square Distribution" before you continue with this handout. That SAS lesson should help you understand the basics of the Chi-Square distribution.

Inferences about Variances and Standard Deviations

Suppose that we have a sample of 31 male high school varsity basketball players. We wish to test the null hypothesis that their heights are a random sample of the general population of men, with respect to variance. We know from selective service records that the standard deviation in the population is 2.5.

Let us use a pair of **directional hypotheses**, which will call for a **one-tailed** test.

$H_{\varnothing}$: $\sigma^2 \geq 6.25$ H_1: $\sigma^2 < 6.25$

We compute the sample variance and find it to be 4.55. We next compute the value of the test statistic, **chi-square**. [If we were repeatedly to sample 31 scores from a normal population with a variance of 6.25 and on each compute $(N-1) * S^2 / 6.25$, we would obtain the chi-square distribution on 30 *df*.]

$\chi^2 = \frac{(df)S^2}{\sigma^2}$ where $df = N-1$ $\chi^2 = 30(4.55) / 6.25 = 21.84$

The expected value of the chi-square (the mean of the sampling distribution) were the null hypothesis true is its degrees of freedom, 30. Our computed chi-square is less than that, but is it enough less than that for us to be confident in rejecting the null hypothesis? We now need to obtain the *p*-value. Since our alternative hypothesis specified a **<** sign, we need find $P(\chi^2 < 21.84 \mid df = 30)$. We go to the chi-square table, which is a one-tailed, upper-tailed, table (in Howell). For 30 *df*, 21.84 falls between 20.60, which marks off the upper 90%, and 24.48, which marks off the upper 75%. Thus, the upper-tailed *p* is $.75 < p < .90$. But we need a lower-tailed *p*, given our alternative hypothesis. To obtain the desired lower-tailed *p*, we simply subtract the upper-tailed *p* from unity, obtaining $.10 < p < .25$. [If you integrate the chi-square distribution you obtain the exact $p = .14$.] Using the traditional .05 criterion, we are unable to reject the null hypothesis.

Our **APA-style summary** reads: "A one-tailed chi-square test indicated that the heights of male high school varsity basketball players ($s^2 = 4.55$) were not significantly less variable than those of the general population of adult men ($\sigma^2 = 6.25$), χ^2 (30, $N = 31$) = 21.84, $p = .14$." I obtained the exact *p* from SAS: "p = **1**-PROBCHI(**127.2**, **100**);". Note that I have specified the variable (height), the subjects (basketball players), the status of the null hypothesis (not rejected), the nature of the test (directional), the parameter of interest (variance),the value of the relevant sample statistic (s^2) the test statistic (χ^2), the degrees of freedom and *N*, the computed value of the test statistic, and an exact *p*. The phrase "not significantly less" implies that I tested directional hypotheses, but I chose to be explicit about having conducted a one-tailed test.

For a **two-tailed** test of **nondirectional** hypotheses, one simply doubles the one-tailed *p*. If the resulting two-tailed *p* comes out above 1.0, as it would if you doubled the upper-tailed *p* from the above problem, then you need to work with the (doubled) *p* from the other tail. For the above problem the two-tailed *p* is $.20 < p < .50$. An APA summary statement would read: A two-tailed chi-square test indicated that the variance of male high school varsity basketball players' heights ($s^2 = 4.55$) was not significantly different from that of the general population of adult men ($\sigma^2 = 6.25$), χ^2 (30, $N = 31$) = 21.84, $p = .28$." Note that with a nonsignificant result my use of the phrase "not significantly different" implies nondirectional hypotheses.

Suppose we were testing the alternative hypothesis that the population variance is greater than 6.25. Assume we have a sample of 101 heights of men who have been diagnosed as having one or more of several types of pituitary dysfunction. The obtained sample variance is 7.95, which differs from 6.25 by the same amount, 1.7, that our previous sample variance, 4.55, did, but in the opposite direction. Given our larger sample size this time, we should expect to have a better chance of rejecting the null hypothesis. Our computed chi-square is 127.2, yielding an (upper-tail) *p* of .025 <

$p < .05$, enabling us to reject the null hypothesis at the .05 level. Our APA-style summary statement reads: "A one-tailed chi-square test indicated that the heights of men with pituitary dysfunction (s^2 = 7.95) were significantly more variable than those of the general population of men ($\sigma^2 = 6.25$), χ^2 (100, N = 101) = 127.2, p = .034." Since I rejected the null hypothesis (a "significant" result), I indicated the direction of the obtained effect ("significantly more variable than ..."). Note that if we had used nondirectional hypotheses our two-tailed p would be $.05 < p < .10$ and we could not reject the null hypothesis with the usual amount of confidence (.05 criterion for α). In that case my APA-style summary statement would read: "A two-tailed chi-square test indicated that the variance in the heights of men with pituitary dysfunction (s^2 = 7.95) was not significantly different from that of the general population of men ($\sigma^2 = 6.25$), χ^2 (100, N = 101) = 127.2, p = .069."

We can also place **confidence limits** on our estimation of a population variance. For a 100(1 $-\alpha$) % confidence interval for the population variance, compute:

$$\left[\frac{(N-1)s^2}{b}, \frac{(N-1)s^2}{a}\right]$$

where a and b are the $\alpha / 2$ and $1 - (\alpha / 2)$ fractiles of the chi-square distribution on $(n - 1)$ df. For example, for our sample of 101 pituitary patients, for a 90% confidence interval, the .05 fractile (the value of chi-square marking off the lower 5%) is 77.93, and the .95 fractile is 124.34. The confidence interval is 100(7.95)/124.34, 100(7.95)/77.93 or 6.39 to 10.20. In other words, we are 90% confident that the population variance is between 6.39 and 10.20. Technically, the interpretation of the confidence coefficient (90%) is this: were we to repeatedly draw random samples and for each construct a 90% confidence interval, 90% of those intervals would indeed include the true value of the estimated parameter (in this case, the population variance).

Please note that the application of chi-square for tests about variances is not robust to the normality assumption made when using such applications. When a statistic is **robust** to violation of one of its assumptions then one can violate that assumption considerably and still have a valid test.

Chi-Square Approximation of the Binomial Distribution

$\chi_1^2 = \frac{(Y-\mu)^2}{\sigma^2}$ where Y is from a normal population.

Consider Y = # of successes in a binomial experiment. With $np \pm 2\sqrt{npq}$ within $0 \to N$, the binomial distribution should be approximately normal. Thus,

$\chi_1^2 = \frac{(Y-np)^2}{npq}$, which can be shown to equal $\frac{(Y-np)^2}{np} + \frac{(n-Y-nq)^2}{nq}$. Here is a proof (the "not-so-obvious algebra" referred to by Howell):

$(n-Y-nq)^2 = [n-Y-n(1-p)]^2 = (n-Y-n+np)^2 = (np-Y)^2$. Now, since

$(a-b)^2 = (b-a)^2, (np-X)^2 = (X-np)^2$.

Thus, $\frac{(Y-np)^2}{np} + \frac{(n-Y-nq)^2}{nq} = \frac{(Y-np)^2}{np} + \frac{(Y-np)^2}{nq} = \frac{q(Y-np)^2 + p(Y-np)^2}{npq}$

$= \frac{(q+p)(X-np)^2}{npq}$, which $= \frac{(Y-np)^2}{npq}$, since $q + p = 1$.

Substituting O_1 for number of successes, O_2 for number of failures, and E for np,

$\chi^2 = \frac{(O_1-E_1)^2}{E_1} + \frac{(O_2-E_2)^2}{E_2} = \sum \frac{(O-E)^2}{E}$

The Correction for Continuity (Yates Correction) When Using Chi-square to Approximate a Binomial Probability

Suppose that we wish to test the null hypothesis that 50% of ECU students favor tuition increases to fund the acquisition of additional computers for student use at ECU. The data are: in a random sample of three, not a single person favors the increase. The null hypothesis is that binomial $p = .50$. The two-tailed exact significance level (using the multiplication rule of probability) is $2 \times .5^3 = .25$.

Using the chi-square distribution to approximate this binomial probability,

$$\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{(0-1.5)^2}{1.5} + \frac{(3-1.5)^2}{1.5} = 3.00$$, $p = .0833$, not a very good approximation.

Remember that a one-tailed p is appropriate for nondirectional hypotheses with this test, since the computed chi-square increases with increasing (O - E) regardless of whether O > E or O < E.

Using the chi-square distribution with Yates' correction for continuity.:

$$\chi^2 = \sum \frac{(|O-E|-.5)^2}{E} = 2\left[\frac{(1.5-.5)^2}{1.5}\right] = 1.33$$, $p = .25$, a much better approximation.

Half-Tailed Tests

Suppose that you wanted to test directional hypotheses, with the alternative hypothesis being that fewer than 50% of ECU students favor the increased tuition? For the binomial p you would simply not double the one tailed $P(Y \leq 0)$. For a directional chi-square, with the direction correctly predicted in the alternative hypothesis, you take the one-tailed p that is appropriate for a nondirectional test and divide it by the number of possible orderings of the categorical frequencies. For this problem, we could have had more favor than disfavor or more disfavor than favor, two possible orderings. This is really just an application of the multiplication rule of probability. One one-tailed p_1 gives you the conditional probability of obtaining results as or more discrepant with the null than are those you obtained. The probability of correctly guessing the direction of the outcome, p_2, is ½. The joint probability of getting results as unusual as those you obtained AND in the predicted direction is $p_1 p_2$.

One-Sixth Tailed Tests

What if there were three categories, favor, disfavor, and don't care, and you correctly predicted that the greatest number of students would disfavor, the next greatest number would not care, and the smallest number would favor? [The null hypothesis from which you would compute expected frequencies would be that 1/3 favor, 1/3 disfavor, and 1/3 don't care.] In that case you would divide your one-tailed p by 3! = 6, since there are 6 possible orderings of three things.

The basic logic of the half-tailed and sixth-tailed tests presented here was outlined by David Howell in the fourth edition of his *Statistical Methods for Psychology* text, page 155. It can be generalized to other situations, for example, a one-way ANOVA where one predicts a particular ordering of the group means.

Multicategory One-Way Chi Square

Suppose we wish to test the null hypothesis that Karl Wuensch gives twice as many C's as B's, twice as many B's as A's, just as many D's as B's, and just as many F's as A's in his undergraduate statistics classes. We decide on a nondirectional test using a .05 criterion of significance. The observed frequencies are: A: 6, B: 24, C: 50, D: 10, F: 10. Under this null hypothesis, given a total N of 100, the expected frequencies are: 10, 20, 40, 20, 10, and $\chi^2 = 1.6 + 0.8 + 2.5 + 5 + 0 = 9.9$; $df = K - 1 = 4$. $p = .042$. We reject the null hypothesis.

There are additional analyses you could do to determine which parts of the null hypothesis are (significantly) wrong. For example, under the null hypotheses one expects that 10% of the grades will be A's. Six A's were observed. You could do a binomial test of the null hypothesis that the proportion of A's is .10. Your two-tailed *p* would be two times the probability of obtaining 6 or fewer A's if $n = 100$ and $p = 0.10$. As an example of another approach, you could test the hypothesis that there are twice as many C's as B's. Restricting your attention to the 50 + 24 = 74 C's and B's, you would expect 2/3(74) = 49.33 C's and 1/3(74) = 24.67 B's. A one *df* Chi-square (or an exact binomial test) could be used to test this part of the omnibus null hypothesis.

Pearson Chi-Square Test for Contingency Tables.

For the dichotomous variables A and B, consider the below joint frequency distribution [joint frequencies in the cells, marginal frequencies in the margins]. Imagine that your experimental units are shoes belonging to members of a commune, that variable A is whether the shoe belongs to a woman or a man, and that variable B is whether the shoe has or has not been chewed by the dog that lives with the commune. One of my graduate students actually had data like these for her 6430 personal data set years ago. The observed cell counts are in bold font.

	A = Gender of Shoe Owner		
B = Chewed?	Female	Male	
Yes	**10** (15)	**20** (15)	30
No	**40** (35)	**30** (35)	70
	50	50	100

We wish to test the null hypothesis that A is independent of (not correlated with) B.

The marginal probabilities of being chewed are .3 chewed, .7 not. The marginal probabilities for gender of the owner are .5, .5.

Using the multiplication rule to find the joint probability of (A = a) $\cap$ (B = b), assuming independence of A and B (the null hypothesis), we obtain .5(.3) = .15 and .5(.7) = .35.

Multiplying each of these joint probabilities by the total *N*, we obtain the expected frequencies, which I have entered in the table in parentheses. A short cut method to get these expected frequencies is: For each cell, multiply the row marginal frequency by the column marginal frequency and then divide by the total table *N*. For example, for the upper left cell, E = 30(50)/100 = 15.

$$\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{(10-15)^2}{15} + \frac{(20-15)^2}{15} + \frac{(40-35)^2}{35} + \frac{(30-35)^2}{35} = 4.762 .$$

Shoes owned by male members of the commune were significantly more likely to be chewed by the dog (40%) than were shoes owned by female members of the commune (20%), $\chi^2(1, N = 100) = 4.762$, $p = .029$, odds ratio = 2.67, 95% CI [1.09, 6.02].

Yates Correction in 2 x 2 Contingency Tables

Don't make this correction unless you find yourself in the situation of having both sets of marginals fixed rather than random. By fixed marginals, I mean that if you were to repeat the data collection the marginal probabilities would be exactly the same. This is almost never the case. There is one circumstance when it would be the case – suppose that you dichotomized two continuous variables using a median split and then ran a 2 x 2 chi-square. On each of the dichotomous variables

each marginal probability would be .5, and that would remain unchanged if you gathered the data a second time.

Small Expected Frequencies

If you have cells in which the expected frequencies are small, your statistical program may warn you that the chi-square analysis may be inappropriate. The major issue here is low power, not inflation of alpha, so I do not worry about this if the result is significant.

Some have advocated use of Fisher's exact test for a 2 x 2 table with small expected frequencies. I am not counted among them, as it assumes that the marginal are fixed, which is highly unlikely.

Ian Campbell has recommended that one use the "*N*-1 Chi-square test" instead of the traditional test when dealing with 2 x 2 contingency tables with small expected frequencies. The value of the *N*-1 Chi-square is most easily calculated as $(N\text{-}1)\phi^2$. Remember that ϕ is simply the Pearson correlation between the (numerically coded) rows variable and columns variable. The regular Chi-square can be calculated as $N\phi^2$.

Misuses of the Pearson Chi-square

Independence of Observations. The observations in a contingency table analyzed with the chi-square statistic are assumed to be independent of one another. If they are not, the chi-square test is not valid. A common way in which this assumption is violated is to count subjects in more than one cell. When I was studying ethology at Miami University I attended a paper session where a graduate student was looking at how lizards move in response to lighting conditions. He had a big terrarium with three environmentally different chambers. Each day he counted how many lizards were in each chamber and he repeated this observation each night. He conducted a Time of Day x Chamber chi-square. Since each lizard was counted more than once, this analysis was invalid. For a 2 x 2 table, McNemar's test may be appropriate when the observations are not independent.

Inclusion of Nonoccurrences. Every subject must be counted once and only once in your contingency table. When dealing with a dichotomous variable, an ignorant researcher might do a one-way analysis, excluding observations at one of the levels of the dichotomous variable. Here is the Permanent Daylight Savings Time Attitude x Rural/Urban example in Howell.

Twenty urban residents and twenty rural residents are asked whether or not they favor making DST permanent, rather than changing to and from it annually: 17 rural residents favoring making DST permanent, 11 urban residents do. An inappropriate analysis is a one-way χ^2 with expected probability of favoring DST the same for rural as for urban residents.

	O	E	$\lvert O\text{-}E\text{-}.5\rvert^2/E$
Rural	17	14	.4464
Urban	11	14	.4464

$\chi^2(1, N = 28) = 0.893, p = .35$

The appropriate analysis would include those who disfavor permanent DST.

	Favor Permanent DST	
Residence	No	Yes
Rural	**3**	**17**
Urban	**9**	**11**

$\chi^2(1, N = 40) = 4.29, p = .038$

See this example of this error in the published literature.

Normality. For the binomial or multinomial distribution to be approximately normal, the sample size must be fairly large. Accordingly, there may be a problem with chi-square tests done with small cell sizes. Your computer program may warn you if many of the expected frequencies are small. You may be able to eliminate small expected frequencies by getting more data, collapsing across (combining) categories, or eliminating a category. Please do note that the primary effect of having small expected frequencies is a reduction in power. If your results are significant in spite of having small expected frequencies, there really is no problem, other than your being less precise when specifying the magnitude of the effect than you would be if you had more data.

Likelihood Ratio Tests

In traditional tests of significance, one obtains a significance level by computing the probability of obtaining results as or more discrepant with the null hypothesis than are those which were obtained. In a likelihood ratio test the approach is a bit different. We obtain two likelihoods: The likelihood of getting the data that we did obtain were the null hypothesis true, and the likelihood of getting the data we got under the exact alternative hypothesis that would make our sample data as likely as possible. For example, if we were testing the null hypothesis that half of the students at ECU are female, $p = .5$, and our sample of 100 students included 65 women, then the alternative hypothesis would be $p = .65$. When the alternative likelihood is much greater than the null likelihood, we reject the null. We shall encounter such tests when we study log linear models next semester, which we shall employ to conduct muldimensional contingency table analysis (where we have more than two categorical variables in our contingency table). See more details here.

Strength of Effect Estimates

I find **phi** an appealing estimate of the magnitude of effect of the relationship between two dichotomous variables and **Cramér's phi** appealing for use with tables where at least one of the variables has more than two levels.

Odds ratios can also be very useful. Consider the results of some of my research on attitudes about animals (Wuensch, K. L., & Poteat, G. M. Evaluating the morality of animal research: Effects of ethical ideology, gender, and purpose. *Journal of Social Behavior and Personality*, 1998, *13,* 139-150. Participants were pretending to be members of a university research ethics committee charged with deciding whether or not to stop a particular piece of animal research which was alleged, by an animal rights group, to be evil. After hearing the evidence and arguments of both sides, 140 female participants decided to stop the research and 60 decided to let it continue. That is, the odds that a female participant would stop the research were 140/60 = 2.33. Among male participants, 47 decided to stop the research and 68 decided to let it continue, for odds of 47/68 = 0.69. The ratio of these two odds is 2.33 / .69 = 3.38. In other words, the women were more than 3 times as likely as the men to decide to stop the research.

Size of effect	$w = \phi$	odds ratio
small	.1	1.49
medium	.3	3.45
large	.5	9

*For a 2 x 2 table with both marginals distributed uniformly.

Why form ratios of odds rather than ratios of probabilities? See my document Odds Ratios and Probability Ratios.

The Cochran-Mantel-Haenzel Statistic

Howell (2013) provides data from a 1973 study of sexual discrimination in graduate admissions at UC Berkeley. In Table 6.8 are data for each of six academic departments. For each department we have the frequencies for a 2 x 2 contingency table, sex/gender of applicant by admissions decision. At the bottom of this table are the data for a contingency table collapsing across departments B through F and excluding data from department A. The data from A were excluded because the relationship between sex and decision differed notably in this department from what it was in the other departments.

The contingency tables for departments B through F are shown below. To the right of each I have typed in the odds ratio showing how much more likely women were to be admitted (compared to men). Notice that none of these differs much from 1 (men and women admitted at the same rates).

The FREQ Procedure

Table 1 of Sex by Decision
Controlling for Dept=B

Sex Decision

Frequency Row Pct	Accept	Reject	Total
F	17 68.00	8 32.00	25
M	353 63.04	207 36.96	560
Total	370	215	585

OR = (17/8)/(353/207)
= 1.25

Table 2 of Sex by Decision
Controlling for Dept=C

Sex Decision

Frequency Row Pct	Accept	Reject	Total
F	202 34.06	391 65.94	593
M	120 36.92	205 63.08	325
Total	322	596	918

OR = 0.88

Table 3 of Sex by Decision
Controlling for Dept=D

Sex Decision

Frequency Row Pct	Accept	Reject	Total	
F	131 34.93	244 65.07	375	OR = 1.09
M	138 33.09	279 66.91	417	
Total	269	523	792	

Table 4 of Sex by Decision
Controlling for Dept=E

Sex Decision

Frequency Row Pct	Accept	Reject	Total	
F	94 23.92	299 76.08	393	OR = 0.82
M	53 27.75	138 72.25	191	
Total	147	437	584	

Table 5 of Sex by Decision
Controlling for Dept=F

Sex Decision

Frequency Row Pct	Accept	Reject	Total	
F	24 7.04	317 92.96	341	OR = 1.21
M	22 5.90	351 94.10	373	
Total	46	668	714	

The CMH statistic is designed to test the hypothesis that there is no relationship between rows and columns when you average across two or more levels of a third variable (departments in this case). As you can see below, the data fit well with that null.

Summary Statistics for Sex by Decision
Controlling for Dept

Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

Statistic	Alternative Hypothesis	DF	Value	Prob
1	Nonzero Correlation	1	0.1250	0.7237
2	Row Mean Scores Differ	1	0.1250	0.7237
3	General Association	1	0.1250	0.7237

Estimates of the Common Relative Risk (Row1/Row2)

Type of Study	Method	Value	95% Confidence Limits	
Case-Control (Odds Ratio)	Mantel-Haenszel	0.9699	0.8185	1.1493
	Logit	0.9689	0.8178	1.1481

The Breslow-Day test is for the null hypothesis that the odds ratios do not differ across levels of the third variable (department). As you can see below, that null is retained here.

Breslow-Day Test for Homogeneity of the Odds Ratios	
Chi-Square	2.5582
DF	4
Pr > ChiSq	0.6342

Total Sample Size = 3593

Below is a contingency table analysis on the aggregated data (collapsed across departments B through F). As you can see, these data indicate that there is significant sex bias again women – the odds of a woman being admitted are significantly less than the odds of a man being admitted.

Sex Decision

Frequency / Row Pct	Accept	Reject	Total	
F	508 28.75	1259 71.25	1767	OR = 0.69
M	686 36.76	1180 63.24	1866	
Total	1194	2439	3633	

Statistics for Table of Sex by Decision

Statistic	DF	Value	Prob
Chi-Square	1	26.4167	<.0001
Likelihood Ratio Chi-Square	1	26.4964	<.0001
Phi Coefficient		-0.0853	

Sample Size = 3633

If we include department A in the analysis, we see that in that department there was considerable sex bias in favor of women.

Table 1 of Sex by Decision
Controlling for Dept=A

Sex	Decision			
Frequency Row Pct	Accept	Reject	Total	
F	89 82.41	19 17.59	108	OR = 2.86
M	512 62.06	313 37.94	825	
Total	601	332	933	

The CMH still falls short of significance, but

The FREQ Procedure

Summary Statistics for Sex by Decision
Controlling for Dept (A through F)

Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

Statistic	Alternative Hypothesis	DF	Value	Prob
1	Nonzero Correlation	1	1.5246	0.2169
2	Row Mean Scores Differ	1	1.5246	0.2169
3	General Association	1	1.5246	0.2169

Estimates of the Common Relative Risk (Row1/Row2)

Type of Study	Method	Value	95% Confidence Limits	
Case-Control (Odds Ratio)	Mantel-Haenszel	1.1053	0.9431	1.2955
	Logit	1.0774	0.9171	1.2658

notice that the Breslow-Day test now tells us that the odds ratios differ significantly across departments. This violates an assumption of the CMH.

Breslow-Day Test for Homogeneity of the Odds Ratios	
Chi-Square	18.8255
DF	5
Pr > ChiSq	0.0021

Total Sample Size = 4526

Here are the data aggregated across departments A through F. Note that these aggregated data indicate significant sex bias against women.

Table of Sex by Decision

Sex	Decision			
Frequency Row Pct	Accept	Reject	Total	
F	597 31.84	1278 68.16	1875	OR = 0.58
M	1198 44.52	1493 55.48	2691	
Total	1795	2771	4566	

Statistics for Table of Sex by Decision

Statistic	DF	Value	Prob
Chi-Square	**1**	**74.4567**	**<.0001**
Likelihood Ratio Chi-Square	1	75.2483	<.0001
Cramer's V		-0.1277	

Howell mentions Simpson's paradox in connection with these data. Simpson's paradox is said to have taken place when the direction of the association between two variables (in this case, sex and admission) is in one direction at each level of a third variable, but when you aggregate the data (collapse across levels of the third variable) the direction of the association changes.

We shall see Simpson's paradox (also known as a reversal paradox) in other contexts later, including ANOVA and multiple regression. See The Reversal Paradox (Simpson's Paradox) and the SAS code used to produce the output above.

Kappa

If you wish to compute a measure of the extent to which two judges agree when making categorical decisions, kappa can be a useful statistic, since it corrects for the spuriously high apparent agreement that otherwise results when marginal probabilities differ from one another considerably.

For example, suppose that each of two persons were observing children at play and at a designated time or interval of time determining whether or not the target child was involved in a fight. Furthermore, if the rater decided a fight was in progress, the target child was classified as being the aggressor or the victim. Consider the following hypothetical data:

	Rater 2			
Rater 1	No Fight	Aggressor	Victim	*marginal*
No Fight	70 (54.75)	3	2	*75*
Aggressor	2	6 (2.08)	5	*13*
Victim	1	7	4 (1.32)	*12*
marginal	*73*	*16*	*11*	*100*

The percentage of agreement here is pretty good, (70 + 6 + 4) ÷ 100 = 80%, but not all is rosy here. The raters have done a pretty good job of agreeing regarding whether the child is fighting or

not, but there is considerable disagreement between the raters with respect to whether the child is the aggressor or the victim.

Jacob Cohen developed a coefficient of agreement, **kappa**, that corrects the percentage of agreement statistic for the tendency to get high values by chance alone when one of the categories is very frequently chosen by both raters. On the main diagonal of the table above I have entered in parentheses the number of agreements that would be expected by chance alone given the marginal totals. Each of these expected frequencies is computed by taking the marginal total for the column the cell is in, multiplying it by the marginal total for the row the cell is in, and then dividing by the total count. For example, for the No Fight-No Fight cell, (73)(75) ÷ 100 = 54.75. Kappa is then computed as: $\kappa = \frac{\sum O - \sum E}{N - \sum E}$, where the *O*'s are observed frequencies on the main diagonal, the *E*'s are expected frequencies on the main diagonal, and *N* is the total count. For our data, $\kappa = \frac{70 + 6 + 4 - 54.75 - 2.08 - 1.32}{100 - 54.75 - 2.08 - 1.32} = \frac{21.85}{41.85} = 0.52$, which is not so impressive.

More impressive would be these data, for which kappa is 0.82:

	Rater 2			
Rater 1	No Fight	Aggressor	Victim	*marginal*
No Fight	70 (52.56)	0	2	*72*
Aggressor	2	13 (2.40)	1	*16*
Victim	1	2	9 (1.44)	*12*
marginal	*73*	*15*	*12*	*100*

Power Analysis

G*Power uses Cohen's *w* as the effect size statistic for contingency table analysis – *w* is nothing more than ϕ. Please read my document Power Analysis for a 2 x 2 Contingency Table .

Nonindependent Observations

Suppose you are evaluating the effectiveness of an intervention which is designed to make patients more likely to comply with their physicians' prescriptions. Prior to introducing the intervention, each of 200 patients is classified as compliant or not. Half (100) are compliant, half (100) are not. After the intervention you reclassify each patient as compliant (120) or not (80). It appears that the intervention has raised compliance from 50% to 65%, but is that increase statistically significant?

McNemar's test is the analysis traditionally used to answer a question like this. To conduct this test you first create a 2 x 2 table where each of the subjects is classified in one cell. In the table below, cell A includes the 45 patients who were compliant at both times, cell B the 55 who were compliant before the intervention but not after the intervention, cell C the 85 who were not compliant prior to the intervention but were after the intervention, and cell D the 15 who were noncompliant at both times.

		After the Intervention		
		Compliant (1)	Not (0)	Marginals
Prior to the Intervention	Compliant (1)	45 A	55 B	100
	Not (0)	85 C	15 D	100
Marginals		130	70	200

McNemar's Chi-square, with a correction for continuity, is computed this way: $\chi^2 = \frac{(|b-c|-1)^2}{b+c}$. For the data above, $\chi^2 = \frac{(|55-85|-1)^2}{55+85} = 6.007$. The chi-square is evaluated on one degree of freedom, yielding, for these data, a *p* value of .01425.

If you wish not to make the correction for continuity, omit the "–1." For these data that would yield a chi-square of 6.429 and a *p* value of .01123. I have not investigated whether or not the correction for continuity provides a better approximation of the exact (binomial) probability or not, but I suspect it does.

For details on how to use SAS, SPSS, or online calculators to conduct McNemar's test, please read my document Comparing Correlated Proportions With McNemar's Test.

Please read the following documents:

- Nondirectional Hypotheses, One-Tailed Test – example with 2 x 2 Pearson chi-square
- Pairwise Comparisons Following a Significant 2 x 3 Contingency Table Analysis
- Congress Chi-Square – Example of contingency table analysis
- Smoking x Drinking – another example
- Contingency Tables with Ordinal Variables – another use of the *N*-1 Chi-Square
- Constructing a Confidence Interval for the Standard Deviation
- Chi-Square, One- and Two-Way -- more detail on the *w* statistic and power analysis
- Power Analysis for One-Sample Test of Variance (Chi-Square)

Return to Wuensch's Stats Lessons Page