Measures of Association for Contingency Tables

The Pearson chi-squared statistic and related significance tests provide only part of the story of contingency table results. Much more can be gleaned from contingency tables than just whether the results are different from what would be expected due to chance (Kline, 2013). For many data sets, the sample size will be large enough that even small departures from expected frequencies will be significant. And, for other data sets, we may have low power to detect significance. We therefore need to know more about the strength of the magnitude of the difference between the groups or the strength of the relationship between the two variables.

Phi

The most common measure of magnitude of effect for two binary variables is the *phi coefficient*. Phi can take on values between -1.0 and 1.0, with 0.0 representing complete independence and -1.0 or 1.0 representing a perfect association. In probability distribution terms, the joint probabilities for the cells will be equal to the product of their respective marginal probabilities, $P(n_{ij}) = P(n_{i+})P(n_{+j})$, only if the two variables are independent. The formula for phi is often given in terms of a shortcut notation for the frequencies in the four cells, called the *fourfold table*.

Azen and Walker Notation

n_{11}	n_{12}
n_{21}	n_{22}

Fourfold table notation

A	B
C	D

The equation for computing phi is a fairly simple function of the cell frequencies, with a cross-multiplication and subtraction of the two sets of diagonal cells in the numerator.[1]

$$\phi = \frac{n_{11}n_{22} - n_{21}n_{12}}{(n_{11}+n_{12})(n_{21}+n_{22})(n_{11}+n_{21})(n_{12}+n_{22})} = \frac{AD - BC}{(A+B)(C+D)(A+C)(B+D)}$$

As you might have expected at this point, the phi coefficient is rather like the correlation coefficient. In fact, it is exactly equal to Pearson's correlation coefficient. The significance test of the correlation coefficient differs slightly from the test of the Pearson χ^2 test, however. The t distribution is used for testing the correlation coefficient, whereas the χ^2 distribution is the square of the z distribution. Of course, the t and z distributions become nearly identical as $n \geq 120$. Though chi-squared is often viewed as a test of homogeneity (group differences) and phi as a measure of association, they are really asking the same question. It is also a simple matter to convert from ϕ to χ^2, where $\chi^2 = n\phi^2$ and $\phi = \sqrt{\chi^2 / n}$. Just as with the correlation coefficient, squaring phi, ϕ^2, gives the proportion of shared variance between the two binary variables. Cramer's V is the generalization of phi for $I \times J$ tables, also simply calculated from chi-square, using the number of levels of whichever is the smaller dimension in the denominator.

$$Cramer's\ V = \sqrt{\frac{\chi^2}{\left[\min(I-1, J-1)\right]n}}$$

Contingency Coefficient

The contingency coefficient, which is often printed in software output but rarely reported by authors, was also suggested by Pearson as a measure of association. The contingency coefficient is intended to estimate the association between two underlying normal variables and can be used for any $I \times J$ table.

[1] The phi coefficient may be represented by the capital Greek letter phi ("fee"), Φ, or the lower case, ϕ. Here, we will represent the sample estimate with a lower case and save the upper case for the population value.

$$Pearson's\ C = \sqrt{\frac{\chi^2}{\left(\chi^2 + n\right)}}$$

Risk and Relative Risk

The concept of *risk* is often used in health research. For example, we may be interested in the risk of coronary heart disease (CHD) for men over 65. If 21 men out of a sample of 100 have heart disease, then the risk is $21/100 = .21$. In the context of contingency table analysis, risk involves marginal frequencies of just one variable, which could be more generally described as $p_{i+} = n_{i+}/n_{++}$ (we could also use columns if desired). The risk difference could be used to compare the risk of two groups. If men have a .21 risk and women have an .11 risk, then the *risk difference* is simply $.21 - .11 = .10$. The two risks can also be compared in the *risk ratio* or *relative risk*, which is $\text{RR} = (n_{2+}/n_{++})/(n_{1+}/n_{++}) = p_{2+}/p_{1+}$. For the hypothetical example below,[2] the risk ratio of CHD for men to women is $.21/.11 = 1.91$.

Hypothetical example of CHD in a sample men and women over 65

	No CHD	CHD	Total
Women	178	22	200
Men	79	21	100

In this example, we might say being male is the presence of a risk for CHD. In general, the relative risk finds the probability of occurrence of the disease when the risk is present compared with when the risk is absent.

$$RR = \frac{P\left(disease \mid risk\right)}{P\left(disease \mid \sim risk\right)}$$

If the relative risk is 1.0, then the probability of occurrence of the disease is the same whether the risk factor is present or absent. But because the RR is 1.91, the probability of CHD is greater for men than women (almost double). Note that we could also frame the question in terms of the relative of risk of women to men, which would be $.11/.21 = .524$, which indicates women have about half the probability of CHD as men.

Standard errors, significance tests, or confidence limits can be computed for risk ratios. The confidence limits for the risk ratio are best stated in terms of the natural log of the risk ratio, $\ln\left(RR\right) \pm \left(1.96\right)\left(SE_{\ln(RR)}\right)$, where

$$SE_{\ln(RR)} = \sqrt{\frac{1 - p_{2+}}{n_{2+}p_{2+}} + \frac{1 - p_{1+}}{n_{1+}p_{1+}}}$$

Odds Ratio

Another way to compare the likelihood of the occurrence of an event between two groups is the *odds ratio*, which is more widely used outside of health and clinical research. The *odds* of an event occurring is the probability of it happening relative to the probability of it not happening. So, in the above example, the odds for men having CHD relative to not having CHD is $21/79 = .266$. That is, men have .266 (about one-quarter) the odds of having CHD relative to not having CHD. The odds by itself is not particularly useful in most circumstances, because we usually want to compare the odds of an event for one group compared with another, which involves two variables rather than just one. The odds ratio is the ratio of odds of the event occurring in one group to the odds of the event occurring in the other group.

[2] These numbers are not entirely arbitrary, as they are based on rates of CHD found in an American Heart Association fact sheet, https://www.heart.org/idc/groups/heart-public/@wcm/@sop/@smd/documents/downloadable/ucm_319574.pdf

$$OR = \frac{odds_2}{odds_1} = \frac{n_{21}/n_{22}}{n_{11}/n_{12}} = \frac{n_{11}n_{22}}{n_{21}n_{12}}$$

The last equation on the right is a quick way to compute the odds ratio by cross-multiplying the two diagonals and dividing. In the fourfold notation, the odds ratio is AD/CB. For the CHD result, the odds of men having CHD relative to women having CHD is $(21/79)/(22/178) = 2.15$, which means that men have more than twice the odds of having CHD as women. An odds ratio of 1.0 means that the odds are equal in the two groups (i.e., there is no relationship between sex and CHD). A positive relationship between the two variables corresponds to odds ratio greater than 1.0 and a negative relationship corresponds to odds ratio less than 1.0. Be careful with interpretation of odds, however. Odds less than 1 should not be interpreted as % chance less of (decrease in) the event occurring. If the odds are .25, it is not a .75 reduction in probability of the events occurrence. It is instead that the second group has four times less the odds of the event occurring, because, if the two groups were switched, the odds ratio would be $1/.25 = 4.0$ instead.

The difference between the risk ratio (relative risk) and the odds ratio can be difficult to grasp. They are different, so do not use the term "risk" when reporting the odds ratio. The risk ratio is asymmetric in that it assumes that the risk precedes the disease causally, whereas the odds ratio does not make the assumption of direction (the rows and columns could be switched and the result would be the same). The odds ratio is also directly tied to logistic regression coefficients (as we will see later, $OR = e^{\beta}$).

$$OR = RR\left(\frac{1-p_{2+}}{1-p_{1+}}\right)$$

For extreme values in which the probability of the event is near zero for both groups, the risk ratio and odds ratio will be similar. When the probability is not near zero, the odds ratio will tend indicate a stronger relationship between the variables (farther away from the value of 1.0 than the risk ratio). Standard errors can be computed for odds ratios for significance testing or confidence limits, but we will save this topic for later when we discuss logistic regression.

Sensitivity and Specificity

Sensitivity and specificity are often useful for clinical applications, such as for describing the accuracy of a diagnosis from a test or examination. Sensitivity represents the probability that a test indicates a patient has a disease (e.g., coronavirus) when the patient truly does have the virus. Specificity then is when the test indicates the patient does not have the disease when the patient truly does not have it. Sensitivity corresponds to the concept of true positive and specificity corresponds to the concept of a true negative. Sensitivity and specificity involve the correct diagnoses, and the incorrect diagnoses can be classified as false positives (the test indicates the patient has the disease when he/she does not, or $1 -$ specificity) or false negatives (failing to diagnose the condition when it is really present, or $1 -$ sensitivity). Using our fourfold table notation, the table below presents the true state of affairs compared to the test result ("case" indicates presence of the disease). Sensitivity and specificity can be computed with the following simple ratios.

		True: Case	True: Non-case
Test result	Case	A	B
	Non-case	C	D

$$Sensitivity = \frac{A}{(A+C)}$$

$$Specificity = \frac{D}{(B+D)}$$

The base rate is a related and valuable quantity to be aware of, defined here simply as the number of true cases $(A + C)$ out of the total $(A + B + C + D)$. Sensitivity and specificity are designed to be independent of the base rate, so no matter how rare or common the disease is, sensitivity and specificity will give accurate estimates of the test's utility in diagnosis. *Positive predictive value*, $A/(A+B)$, and

negative predictive value, $D/(C+D)$, are sometimes used to support the accuracy of the test, but these two measures are sensitive to base rates, with high base rates resulting in poor utility of negative predictive values and low base rates resulting in poor utility of positive predictive values (Glaros & Kline, 1988).

Cohen's Kappa

Kappa (Cohen, 1960) is a measure of agreement, frequently used as a metric of interrater agreement. One can simply count the number of times that Rater 1 is in agreement with Rater 2 on a binary measure (e.g., observing whether a child exhibits an aggressive behavior or not). The straight percentage agreement would tend to overestimate the amount of agreement, because by chance we will expect some agreement between the two raters. Setting up a contingency table with yes/no for one rater on the rows and yes/no for the other rater on columns gives the counts or proportions of agreement along the main diagonal. Summing the diagonal proportions, $\sum_{i=1}^{I} p_{ii}$, gives the overall observed proportion, p_O. The expected proportions are then derived from marginal proportions and represent what is expected by chance, $\sum_{i=1}^{I} p_{i+}p_{+i}$, indicated by p_E. The sample estimate of Kappa is then

$$\kappa = \frac{\sum_{i=1}^{I} p_{ii} - \sum_{i=1}^{I} p_{i+}p_{+i}}{1 - \sum_{i=1}^{I} p_{i+}p_{+i}} = \frac{p_O - p_E}{1 - p_E}$$

Though we can compute a standard error and significance test for kappa, that is typically not of interest, because we want to have a level of agreement that is much higher than chance. The standard for a "good" or "acceptable" kappa value is arbitrary and it depends on the specific area of research and standards of practice. Fleiss' (1971) arbitrary guidelines (<.4 is poor, .4-.75 is fair to good, and > .75 is excellent) seem to be cited most often.

Software Illustrations

I present below the SPSS, R, and SAS syntax for generating many of these measures of agreement using the Quinnipiac survey, but I include only the output from the SPSS run (some or most of them were seen in the previous handout on contingency table tests).

SPSS

```
crosstabs /tables=ind by response
  /cells=count row column total expected
  /statistics=chisq phi risk cc.
```

R

```
> #measures of assocaion
> library(vcd)
> data <- matrix(c(338,363,125,156),ncol=2,byrow=TRUE)
> assocstats(data)
```

SAS

```
proc freq data=one;
        tables ind*response /chisq agree;
run;
```

ind ind vs affil * response intended vote Crosstabulation

			response intended vote		
			0 Trump	1 Biden	Total
ind ind vs affil	.00 affiliate	Count	338	363	701
		Expected Count	330.5	370.5	701.0
		% within ind ind vs affil	48.2%	51.8%	100.0%
		% within response intended vote	73.0%	69.9%	71.4%
		% of Total	34.4%	37.0%	71.4%
	1.00 independent	Count	125	156	281
		Expected Count	132.5	148.5	281.0
		% within ind ind vs affil	44.5%	55.5%	100.0%
		% within response intended vote	27.0%	30.1%	28.6%
		% of Total	12.7%	15.9%	28.6%
Total		Count	463	519	982
		Expected Count	463.0	519.0	982.0
		% within ind ind vs affil	47.1%	52.9%	100.0%
		% within response intended vote	100.0%	100.0%	100.0%
		% of Total	47.1%	52.9%	100.0%

Symmetric Measures

		Value	Approximate Significance
Nominal by Nominal	Phi	.034	.290
	Cramer's V	.034	.290
	Contingency Coefficient	.034	.290
N of Valid Cases		982	

Risk Estimate

	Value	95% Confidence Interval Lower	Upper
Odds Ratio for ind ind vs affil (.00 affiliate / 1.00 independent)	1.162	.880	1.535
For cohort response intended vote = 0 Trump	1.084	.932	1.261
For cohort response intended vote = 1 Biden	.933	.822	1.059
N of Valid Cases	982		

$$RR = (n_{2+}/n_{++})/(n_{1+}/n_{++}) = p_{2+}/p_{1+} = (156/519)/(125/463) = 1.13$$

$$OR = \frac{n_{11}n_{22}}{n_{21}n_{12}} = \frac{(338)(156)}{(125)(363)} = 1.16$$

Sample Write-up

A chi-square test was used to determine whether there was a significant difference between the proportion of Biden and Trump's supporters who are independent. Results indicated that 30.1% of Biden's supporters were independents, whereas 27.0% of Trumps supporters were independents. This difference was not significant, $\chi^2(1) = 1.12$, p = .29 The phi coefficient, $\phi = .03$, suggested a small effect of approximately .3% shared variance. The relative risk ratio, 1.13, indicated that independents were roughly 13% more likely to support Biden than Trump. The odds ratio, 1.16, indicated that the odds of supporting Biden if the respondent was an independent were 1.16 the odds of support for Trump if the respondent was an independent.

References and Further Reading

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*. 37-46.

Fleiss, J.L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76* (5): 378–382.

Glaros, A. G., & Kline, R. B. (1988). Understanding the accuracy of tests with cutting scores: The sensitivity, specificity, and predictive value model. *Journal of Clinical Psychology, 44*(6), 1013-1023.

Kline, R.B. (2013). *In Beyond significance testing: Statistics reform in the behavioral sciences*. (2nd ed.). Washington, DC: American Psychological Association