Examining the Computation of the Chi-square Statistical Test

By Shauna M. Overgaard, MHI

Outcome measures such as patient satisfaction, rates of readmission, and procedural complications play heavily into the penalization of an organization's financial reimbursement.

It is important to evaluate data on multiple levels in order to consider all opportunities for intervention. In the case of surgical procedural complications, we might consider compliance, geographic location, and surgical volume. It is becoming increasingly important for health information management professionals to consider possible variables that may affect important outcomes. Whether a fan of statistics or not, because health information management (HIM) professionals have the opportunity to delve into contributing factors, it is important for those analyzing or interpreting the data to have a foundational understanding of how these tests are computed and how they can help translate data into useful information.

In a *Data Revolution* post in August, we described analytical tools to determine the positive and negative predictive values of a test and applied the underlying formulas to real data. As part of an effort to support HIM professionals, we now provide a real-world example of the chi-square test. Based on the problem being evaluated, the chi-square test determines whether a relationship exists between categorical variables (i.e., sex, discharge status, admission source, MS-DRG, etc.). We'll begin by explaining the underlying formula.

Underlying Formula for Chisquare Test

When our null hypothesis is true, the probability of being diseased and exposed is equal to the probability of being diseased multiplied by the probability of exposure.

Null Hypothesis True

The expected cell frequency is equal to the sum of the row in which the cell is found, multiplied by the column in which that cell is found, divided by the sum of all cells combined. Or perhaps even more simply, using AHIMA author Susan White's equation found in A Practical Approach to Analyzing Heathcare Data, we may say, "row total multiplied by column total divided by the grand total of all cells."
Expected Cell Frequency Equation

That is, under the null hypothesis, we ought to "expect" (*E*) the cell count in the top left cell of the table to be:

Expected Cell Count Equation

We can apply this equation to determine the expected value of cells b, c, and d.

Again, to find the expected count of cell b, we multiply the row total (a + b) by the column total (b + d) and divide by the grand total of all cells (n).

Expected Cell Count Equations Continued

For each cell we now have both expected counts (E_a, E_b, E_c, E_d) and observed counts (a, b, c, d). From these values we can determine how near the observed counts are to what we expected to see.

It is from these pieces that we can construct the chi-square (X^2) statistic, according to lecture notes distributed by Wolfson at University of Minnesota:

Chisquare Equation

When the null hypothesis (H_0: the variables are independent) is true, we would expect most of the observed values (O_i) to be very close to the corresponding expected (E_i) values, so $(O_i – E_i)^2$ will be small and will be near zero.

When H_0 is false, we would expect one or more of the O_i values to be far from the corresponding E_i values, so $(0_i – E_i)^2$ will be large and will be large.

Now, this test statistic (X^2) must always be positive and we will choose to reject the null hypothesis (H_0) when the chi-square tests statistic is large.

In order to breathe life into our table and equations, let's use real data from a recently published study. Consider this subsection of the overall analysis that refers to hospital characteristics and the rate of complication experienced by surgical patients. Specifically, is surgical complication rate associated with surgical volume?

We can begin by creating our contingency table using the observed data found in Table 1 of Tevis' full-text paper:
Contingency Table

Our null hypothesis is that complication rate and surgical volume are independent.

Null Hypothesis Example

The alternative hypothesis is that complication rate and surgical volume are dependent:

Alternative Hypothesis Example

We can bring back our expected cell frequency formula from White's text:

Expected Cell Frequency Example

It's important to note that the column and row totals are not altered in the calculation of expected rates. Rather, the proportion of frequencies within the cells have been redistributed.

To continue in our calculation, we may now use these expected values and calculate the chi-square test statistic by inserting them, along with our observed values, into the equation.

Chisquare Example

Now we can take the test statistic and compare it to the critical value that is based on the chi-square distribution. In order to conclude statistical significance, the test statistic must be larger than the critical value. The chi-square distribution depends on a value called "degrees of freedom." Each critical value has a mathematically pre-determined corresponding P-value for a given number of variables that are compared. In order to evaluate this, we first find the "degrees of freedom" for our data set. To explain further, the more variables there are in a calculation, the more likely it is you will find significance in one of them purely by chance. The degrees of freedom can be calculated by computing the product of the number of rows minus 1 and the number of columns minus 1.

Degrees of Freedom Calculation

In our case, the contingency table we are working with has four cells of original data; the totals and headings don't count in this equation. That is, we have two rows and two columns. Using the equation we compute the degrees of freedom and determine that our distribution has one degree of freedom.

Degrees of Freedom Calculation Example

For each value of "degrees of freedom" there is a corresponding critical values table that matches the chi-squared distribution with a corresponding P-value. These tables are widely available on the web.

Degrees of Freedom Values and Pvalues

While our preset alpha was only 0.05, we can see that the value of our test statistic far exceeds the critical value of significance, p<0.001. As described in the paper, given the data we have and the test we've run it appears that the rate of complications is significantly dependent on the hospital surgical volume. At first glance this may seem counterintuitive as the high volume centers may be known for their expertise and we might suspect that high-throughput centers would be running well-established operating procedures. This hints at a potential confound, the difficulty level of the case. Thus, it may be wise to consider additional contributing factors. For example, challenging cases may be sent to high volume centers and the complications associated may be contingent on the difficulty level of the case.

References

Overgaard, Shauna M. and Julie A. Dooling. "Measurement Brings Meaning to Population Health. *Journal of AHIMA* website. August 27, 2015. http://journal.ahima.org/2015/08/27/measurement-brings-meaning-to-population-health/.

White, Susan. "Analyzing Categorical Variables." *A Practical Approach to Analyzing Healthcare Data*, Second Edition. Chicago: AHIMA Press, 2013. pp. 55-76.

Wolfson, J. "Case Control Studies." Lecture notes, distributed in Public Health 6450, Biostatistics I, at The University of Minnesota. March 27, 2013.

Shauna Overgaard, MHI, (sovergaard@css.edu) is an adjunct professor of healthcare data analytics in the Department of Health Informatics and Information Management, for The College of St. Scholastica's CAHIIM accredited MS program. She is a PhD student of Biomedical Health Informatics and Biostatistics at the University of Minnesota.

Original source:
Overgaard, Shauna M. "Examining the Computation of the Chi-square Statistical Test" (Journal of AHIMA website), October 2015.
