Handout 13: Chi-Squared Test

There is a popular alternative to the log-likelihood G-scores that we have been using for testing the goodness-of-fit of discrete data and checking the independence of columns and rows in a contingency table called the **chi-squared test**. You might be thinking one (or both) of the following: (1) haven't we already learned the chi-squared test when testing the variance of one sample? or (2) does the G-score have a chi-squared distribution? You would be correct in both of these points and in being somewhat confused. Some of the original hypothesis tests generated in the history of statistics unfortunately became known by the name of the null-distribution, which is very confusing because it later turned out that those distributions had many other applications. At any rate, the test that most people would mean when the call something **the** chi-squared test, is the one I will show you today.

The chi-squared test starts with the same setup as the log-likelihood test with G-scores, namely, we compute the expected counts e_i or $e_{i,j}$ for each cell in the table. The form of the test statistic is different, but should look somewhat similar (I will write it for the multinomial case, but it should be clear that it is the same idea with the contingency table):

$$C = \sum_i \frac{(x_i - e_i)^2}{e_i} \sim \chi^2.$$

The degrees of freedom of the chi-squared test will be exactly the same as with the G-scores: the number of cells minus one, minus the number of free parameters in the null-hypothesis.

These two tests are actually related to one another. Let $\Delta_i = x_i - e_i$. Then, starting with the G-score we have:

$$\begin{aligned} G &= -2\sum_i x_i \log\left(\frac{e_i}{x_i}\right) \\ &= 2\sum_i x_i \log\left(\frac{x_i}{e_i}\right) \\ &= 2\sum_i (e_i + \Delta_i) \log\left(\frac{e_i + \Delta_i}{e_i}\right) \\ &= 2\sum_i (e_i + \Delta_i) \log\left(1 + \frac{\Delta_i}{e_i}\right) \end{aligned}$$

Now, we can use the Taylor series of $\log(y+1)$ around 1, which gives (for small y): $y - 0.5y^2 + O(y^3)$.[1] Then:

[1] The notation $O(y^3)$ means a polynomial of y that has no constant, linear, or quadratic terms. When y is small, those other terms will dominate the higher order ones, so we do not worry about keeping track of them.

$$
\begin{aligned}
G &= 2\sum_i (e_i + \Delta_i)\left(\frac{\Delta_i}{e_i} - \frac{\Delta_i^2}{2e_i^2} + O(\Delta_i^3)\right) \\
&= 2\sum_i \left[\Delta_i - \frac{\Delta_i^2}{2e_i} + \frac{\Delta_i^2}{e_i} - \frac{\Delta_i^3}{2e_i^2} + O(\Delta_i^3)\right] \\
&= 2\sum_i \left[\Delta_i + \frac{\Delta_i^2}{2e_i} + O(\Delta_i^3)\right]
\end{aligned}
$$

We know that $\sum_i \Delta_i = 0$, because the expected counts and actual counts both sum to 1. So, after removing the higher-order terms:

$$
G \approx \sum_i \frac{\Delta_i^2}{e_i} = \sum_i \frac{(x_i - e_i)^2}{e_i} = C.
$$

So, under the null hypothesis in which we do not expect the differences between the observed and expected counts to differ too much, the two tests can be seen as approximations of one another.

The chi-squared test, as implied by being named by the corresponding distribution, is very old and has been ingrained as the *go-to* test for contingency tables in some domains. I would say that most statisticians would now recommend the G-test as the best alternative, but you will see the chi-squared test in many sources and should understand it and know how to use it and how it compares to the likelihood-ratio test.

What's better about the G-test in most situations? Both tests are asymptotically valid under the null hypothesis, though the G-test is more robust to small sample sizes. More importantly, the chi-squared test can only be motived under the null hypothesis. It's hard to say what is measures, if anything, under an alternative hypothesis. That's not true of likelihood-ratio tests: the value of Λ (which we call G in this specific sub-case) is a reasonable measurement of *how much* the MLE-based model of the data differs from the null hypothesis. So, for the kinds of analysis we were doing last class, the G-scores tend to be much better for finding those relationships which differ the most.