Handout 12: Contingency Tables

While there are many uses of the G-test, the most common application is in the study of contingency tables. Consider, for example, a multinomial with $k = 4$, just as before. However, this time we are going to arrange the data into a two-by-two table, using a slightly different notation for the counts to make it clear that each is associated with a specific row and column. This yields the following, where we have added row sums r_j and column sums c_j, since we will need them in a moment:

$x_{1,1}$	$x_{1,2}$	r_1
$x_{2,1}$	$x_{2,2}$	r_2
c_1	c_2	n

We can re-define the multinomial probabilities similarly, where $p_{i,j}$ is the probability of landing in row i and column j. A very common type of hypothesis test is to consider the set Θ_0 of all tables in which event of being in row i is independent of the event of being in column j, for all combinations of i and j.

The maximum likelihood estimator is unchanged in this case; it is still the raw counts divided by the sample size. The numerator of the G test, however, is different. In order to be in Θ_0, we need to have that $p_{i,j}$ is equal to the probability of being in row i times the probability of being in column j. It should not be surprising to know then that in order to maximize the log-likelihood under H_0, we use the following probabilities and implied expected counts:

$$\tilde{p}_{i,j} = \left(\frac{r_i}{n}\right) \times \left(\frac{c_j}{n}\right) \quad \Rightarrow \quad e_{i,j} = \left(\frac{r_i \times c_j}{n}\right).$$

In other words, the proportion of data that were in row i times the proportion of data that were in column j. From here, we use the same formula as we have on the other page by replacing the sum of j with a double sum over both i and j.

We can extend this same approach to the case where we have R rows and C columns. What, in general, will be the degrees of freedom for G? We have $CR - 1$ dimensions in Θ (any set of probabilities, with the one restriction that the sum to 1) and $(C-1) + (R-1)$ in Θ_0 (any set of valid column probabilities and row probabilities, each having to sum to 1). This difference factors as:

$$(CR - 1) - (C - 1) - (R - 1) = (C - 1) \cdot (R - 1).$$

So, in the common two-by-two table case, we have only a single degree of freedom. This will grow larger for tables with more rows and/or columns.