Handout 11: Multinomial Distribution

Recall that the binomial distribution can be thought of as doing n flips of coin that lands heads with probability p and counting the number of resulting heads. The **multinomial distribution** is a generalization of this that can be conceptualized as rolling a k-sided die n times and counting the number of times that it lands on each side. As usual, one of the hardest things is picking a good notation. Let $x_1, \ldots, x_k$ represent a specific set of counts (these are integers that sum up to n) and $p_1, \ldots, p_k$ be the probabilities of landing on each side (positive values that sum up to 1). Using the counting theorems from probability theory, we can see that the likelihood function will have the following form:

$$\mathcal{L}(p_1, \ldots, p_k; x_1, \ldots, x_k) = \frac{n!}{x_1! \cdots x_k} \times p_1^{x_1} \cdots p_k^{x_k} = \frac{n!}{\prod_j x_j} \times \prod_j p_j^{x_j}.$$

The multinomial is very useful in statistics because we can use it to model any distribution over a set of categories. The MLE estimators for the p_j's has a very nice form:[1]

$$\hat{p}_j = \frac{x_j}{n}, \quad j \in \{1, \ldots, k\}.$$

This just says that our best guess for the probability of being in category j is equal to the proportion of the data that was observed in category j.

The interesting thing happens when we consider the likelihood-ratio test for multinomial data. Let's consider testing the null hypothesis that the true probabilities are $\tilde{p}_1, \ldots, \tilde{p}_k$.[2] This gives, since we already know the MLE, the following value for G:

$$G = -2 \cdot \log \left[\frac{\frac{n!}{\prod_j x_j} \times \prod_j \tilde{p}_j^{x_j}}{\frac{n!}{\prod_j x_j} \times \prod_j \hat{p}_j^{x_j}} \right] = -2 \cdot \log \left[\prod_j \left(\frac{\tilde{p}}{\hat{p}} \right)^{x_j} \right]$$

Now, let's convert the null-hypothesis from probabilities into expected counts: $e_i = \tilde{p} \cdot n$. Also plugging in the form of the MLE, we then have:

$$G = -2 \cdot \sum_j x_j \cdot \log(e_j / x_j).$$

This specific application of the log-likelihood ratio test is often called the **G-test**, hence the reason I have used this letter throughout to refer to this test statistic.

[1] The derivation is not too tricky, but requires using a constrained optimization technique such as Lagranian multipliers, which I do not think everyone has seen. The result is very intuitive, so we will skip the proof.

[2] Our usual notion used a zero in the subscript for the parameters of the null-hypothesis, but we already have subscripts for the different probabilities, which is why I am using a tilde instead.

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.