Worksheet 05: Contingency Tables

Today we want to consider testing how consistent the counts of a categorical variable are with a specific distribution over the categories. We will work through a small example where the derivation is most tracktable, and then extend to the more general cases.

Consider n independent observations of a distribution that can take on one of two values. We will use the random variables O_1 and O_2 to indicate the number of observations in the first and second categories, respecively. Assume that we want to do an hypothesis test where H_0 assume that the true probability of being in the first category is some constant p_1; the proabability of being in the second group will then be $p_2 = 1 - p_1$. A helpful intermediate step is to convert these probabilities into expected counts E_i. That is, $E_1 = p_1 \cdot n$ and $E_2 = p_2 \cdot n$. Our goal is to find a test statistic with a known distribution under this null hypothesis.

There are several options for this case with just two categories. **Q01.** What is the distribution of O_1 under the null hypothesis? What are the expected value and variance of O_1 (you can look these up on the table of distributions). This has a $Bin(n, p_1)$ distribution, so its mean is np_1 and its variance is $np_1(1 - p_1)$.

The distribution of O_1 indicates one approach for an hypothesis test. **Q02.** Write down a test statistic Z that involves only O_1 and E_1 that has an asymptotically standard normal distrbution. The difference $O_1 - E_1$ will have an expected value of 0 under H_0 and it will have a variance equal to $np_1(1 - p_1)$. So, a test statistic that we could use is:

$$Z = \frac{O_1 - E_1}{\sqrt{np_1(1 - p_1)}}$$

Which will limit to a standard normal as the sample size increases.

The test statistic from the previous question is a great (and perhaps best) specific example. However, it cannot be easily extended to the case with more than two categories. As an alternative, consider the following test statistic:

$$C = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2}$$

Plugging in the values of E_i in terms of p_1 and n and the value of O_2 in terms of O_1, we see that:

$$\begin{aligned}
C &= \frac{(O_1 - np_1)^2}{np_1} + \frac{(n - O_1 - n(1 - p_1))^2}{n(1 - p_1)} \\
&= \frac{(O_1 - np_1)^2}{np_1} + \frac{(O_1 - np_1)^2}{n(1 - p_1)} \\
&= \left(\frac{O_1 - np_1}{\sqrt{n}}\right)^2 \cdot \left[\frac{1}{p_1} + \frac{1}{1 - p_1}\right] \\
&= \left(\frac{O_1 - np_1}{\sqrt{n}}\right)^2 \cdot \left[\frac{1 - p_1 + p_1}{p_1(1 - p_1)}\right] \\
&= \left(\frac{O_1 - np_1}{\sqrt{np_1(1 - p_1)}}\right)^2
\end{aligned}$$

Q03. What is the asymptotic distribution of C? The quantity inside of the parenthesis is a binomial minus its expected value and divided by the square root of its variance. So, this will asymptotically be equal to a standard normal. We are squaring this quantity, so we have $C \sim \chi_2(1)$, a chi-squared with one degree of freedom.

We can extend the result with two categories where we have n observations of a categorical variable from m different categories. Let $O_1, \ldots, O_m$ be the observed counts from m categories. Under the assumption that the expected counts should be $E_1, \ldots, E_m$, respecively, define the following test statistic:

$$C = \sum_{i=1}^{m} \frac{(O_i - E_i)^2}{E_i}$$

Then, in the limit of a large sample size, C will have a $\chi_2(m-1)$ distribution. We will not derive this distribution as the mathematics gets quite messy, but you hopefully get the intuition based on the two-sample case.

Let's try to apply this technique to some data. Consider rolling a six-sided die 120 times. We get the following counts of the six different numbers:

$$O_1 = 21, O_2 = 22, O_3 = 16, O_4 = 17, O_5 = 19, O_6 = 25.$$

Q04. Consider an hypothesis test where H_0 is that the die is *fair*; that is, each side is equally likely. What are the values of E_i? Each of the E_i should be 20, the sample size multiplied by 1/6.

Q05. What is the value of the test statistic C for this data? How many degrees of freedom does the chi-squared distribution have under H_0? We have:

$$\begin{aligned} C &= \frac{(21-20)^2}{20} + \frac{(22-20)^2}{20} + \frac{(16-20)^2}{20} + \frac{(17-20)^2}{20} + \frac{(19-20)^2}{20} + \frac{(25-20)^2}{20} \\ &= 2.8 \end{aligned}$$

Note that the quantity is easier to calculate on a calculator if you do the division by 20 only once at the end. The chi-squared has 5 degrees of freedom, one less than the number of categories. You can use R to find the exact p-value. The (one-sided) p-value is 0.7307.

The example above is a form of a **goodness-of-fit** test, an hypothesis test to check if a set of data follows a specific distribution. We will see more of these throughout the semester to deal with continuous random variables. Another usage of the chi-squared test is to do a test of independence between two categorical variables. Let's look at some example data from a sample of 500 adults is regarding their political affiliation and opinion on a tax reform bill. Here is the data as a 2-by-3 table:

	favor	indifferent	opposed	total
democrat	138	83	64	285
republican	64	67	84	215
total	202	150	148	500

We have added the row and column totals as they will be needed in the next step, but note that those counts are external to the raw data of the six counts. We want to do an hypothesis test where H_0 is that the row variable and the column variable are independent. We cannot use the exact same chi-squared test from the six-sided die example because we don't have specific expected counts for each of the six measurements under the null hypothesis. The solution to this is to first convert the totals into percentages and remove the raw data counts:

	favor	indifferent	opposed	total
democrat				0.570
republican				0.424
total	0.404	0.300	0.296	1.000

Recall from probability theory that under the assumption that the rows and columns are independent, the probability of every cell should be just the product of the two probabilities. So, we have the following probabilities under H_0:

	favor	indifferent	opposed	total
democrat	0.230	0.171	0.169	0.570
republican	0.171	0.127	0.126	0.424
total	0.404	0.300	0.296	1.000

And then, we can get the expected counts by multiplying each of these by 500:

	favor	indifferent	opposed	total
democrat	115.0	85.5	84.5	
republican	85.5	63.5	63.0	
total				

Now, we have expected counts under H_0 and can compute a test statistic C just as before using the expected and observed counts:

$$\begin{aligned} C &= \frac{(138-115.0)^2}{115.0} + \frac{(83-85.5)^2}{85.5} + \frac{(64-84.5)^2}{84.5} + \frac{(64-85.5)^2}{85.5} + \frac{(67-63.5)^2}{63.5} + \frac{(84-63.0)^2}{63.0} \\ &= 22.24 \end{aligned}$$

For the same reasons as the previous case, C will have an approximately chi-squared distribution. The degrees of freedom will be the number of rows minus one times the number of columns minus one. Why? We needed to use the data to estimate the column and row proportions, though the final proportion is known after we have the others. The p-value is quite small (≤ 0.001), indicating that we should reject the null hypothesis that political affiliation is independent from their opinion on the tax bill.