Basic Introduction to Contingency Tables

Gene Katsevich, Kenneth Tay *

July 2, 2018

1 What is a contingency table?

A contingency table is simply a tabulation of the empirical joint distribution of a set of categorical random variables.

Let's say that $(X, Y) \in [I] \times [J]$ are categorical random variables (here $[I]$ denotes $\{1, \dots, I\}$). Suppose that $\mathbb{P}[X = i, Y = j] = \pi_{ij}$. We observe i.i.d. observations (X_m, Y_m) for $m = 1, \dots, n$, and we tabulate

$$n_{ij} = \#\{m : X_m = i, Y_m = j\}.$$

If we take n_{ij} to be the (i, j) entry of a neat little grid, we have ourselves an $I \times J$ contingency table!

We may have higher order contingency tables if we want to study more random variables. For example, we might have $(X, Y, Z) \in [I] \times [J] \times [K]$. We can also think of the counts for one categorical random variable as a contingency table of order 1.

2 Sampling models for contingency tables

For simplicity, let's consider 2×2 tables, whose entries are $(n_{11}, n_{12}, n_{21}, n_{22})$.

There are several sampling models for this table, characterized by what margins of the table (if any) we condition on. These sampling models are what we use to get estimates for our parameters, or to construct hypothesis tests.

*light edits by Stephen Bates

No conditioning: Poisson sampling. The unconditional model for 2×2 tables is

$$n_{ij} \overset{\text{ind}}{\sim} \text{Poi}(\mu_{ij}) = \text{Poi}(\mu\pi_{ij}), \tag{1}$$

where $\sum_{ij} \pi_{ij} = 1$. This models the scenario when we did not fix any sample sizes in advance, and just observed a certain process with these four possible outcomes for a certain period of time.

Conditioning on table total: multinomial sampling. Suppose we collected a total of n people, and then observed for each of them whether they smoked and whether they have lung cancer. Then, we are conditioning on the total number of entries in the table, which results in the multinomial sampling model:

$$n_{ij} \sim \text{Mult}(n, \pi_{ij}).$$

Conditioning on row totals: binomial sampling. Consider a prospective study, where we fixed n_{1+} and n_{2+}, the number of people assigned to the treatment and control groups. Then we observed an outcome for each person. In this case, the rows of the table are distributed as independent binomials:

$$n_{11} \sim \text{Bin}(n_{1+}, \pi_1), \quad n_{21} \sim \text{Bin}(n_{2+}, \pi_2), \quad \pi_1 = \frac{\pi_{11}}{\pi_{11} + \pi_{12}}, \ \pi_2 = \frac{\pi_{21}}{\pi_{21} + \pi_{22}}.$$

Conditioning on column totals: binomial sampling. Consider a retrospective case-control study, where we fixed n_{+1} and n_{+2}, the number of people with a disease and the number of people without a disease. For each person, we measured whether they had a certain exposure (e.g. smoking). In this case, the columns of the table are distributed as independent binomials:

$$n_{11} \sim \text{Bin}(n_{+1}, \gamma_1), \quad n_{12} \sim \text{Bin}(n_{+2}, \gamma_2), \quad \gamma_1 = \frac{\pi_{11}}{\pi_{11} + \pi_{21}}, \gamma_2 = \frac{\pi_{12}}{\pi_{12} + \pi_{22}}.$$

Conditioning on row and column totals: hypergeometric sampling. Consider conditioning on the row and column totals of the table. There is one degree of freedom left in the table: once n_{11} is fixed, the rest of the table is determined.

3 What questions are we trying to answer?

Here are the 3 most basic questions we might want to answer for contingency tables:

1. **Testing for independence.** Perhaps the most common question to ask about an $I \times J$ table is whether X and Y are independent. Viewing Y as a response and X as an explanatory variable, independence implies that the distribution of Y is the same no matter what X is, e.g. the chances of getting a heart attack are the same regardless of whether you take aspirin or a placebo.

2. **Testing for goodness of fit.** Here, we want to test if the binned counts we see match some theoretical distribution. This is most commonly used when we have binned counts for a single continuous random variable, and we want to test whether it came from some distribution (e.g. Exp(1)).

For each of these cases, we can use the **Pearson and generalized likelihood ratio tests**. The strategy is as follows:

1. For each cell (i, j), compute the expected count for the cell under the null hypothesis (which we denote by $\hat{\mu}_{ij}$). The expected counts are computed from the MLE fit on the whole table. In the special case of testing independence, the expected counts then depend only on the marginal totals.

2. Compute either the Pearson χ^2 statistic:

$$X^2 = \sum_{i,j} \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}},$$

   or the likelihood ratio test statistic:

$$G^2 = -2 \log \Lambda = 2 \sum_{i,j} n_{ij} \log(n_{ij}/\hat{\mu}_{ij}),$$

   where Λ is the likelihood ratio.

3. Reject if the test statistic is large. Both of these statistics have asymptotic χ^2 distributions, with the degrees of freedom depending on the test we are running.

   - For test of independence, $df = (I - 1)(J - 1)$.
   - For test of goodness of fit, if we estimate d parameters under the null with IJ total entries in the table, $df = IJ - 1 - d$.

Fisher's exact test for 2×2 tables

For testing independence in 2×2 tables, we have another option. The tests above rely on asymptotic χ^2 distributions; If we have a small sample, then the asymptotics might not have kicked in yet. In the 2×2 case, we can use Fisher's exact test instead.

Under the null hypothesis of independence, conditional on the row and column margins of a 2×2 table, $n_{11} \sim \text{HyperGeom}(n_{1+}, n_{+1}, n)$. Hence, we use this exact null distribution to compute small-sample p-values. This is called **Fisher's exact test** or the hypergeometric test.

4 Loglinear and logistic regression models

Loglinear regression (sometimes called Poisson regression) and logistic regression models are the big guns you bring out for serious contingency table analysis. Basically all of the tests above are just special cases of various tests of parameters in these GLMs. Logistic regression models capture conditional distributions $Y|X$, so are appropriate for cases when Y is viewed as a response and X is explanatory. On the other hand, loglinear models capture the joint distribution of (X, Y), so are appropriate when we want to view all variables as response variables. Nevertheless, the two kinds of models have strong connections and can often be translated into each other.

Loglinear models for independence and interaction in $I \times J$ tables.

- Independence is modeled
$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y.$$
For identifiability, we'll need constraints such as $\lambda_I^X = \lambda_J^Y = 0$ or $\sum_i \lambda_i^X = \sum_j \lambda_j^Y = 0$.

- The saturated model is
$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY}.$$
The λ_{ij}^{XY}'s are association terms which reflect deviations from independence. For identifiability, with constraints $\lambda_I^X = \lambda_J^Y = 0$, we'll need further constraints $\lambda_{Ij}^{XY} = \lambda_{iJ}^{XY} = 0$. There are exactly IJ parameters in this model, and the ML fitted values are $\{\hat{\mu}_{ij} = n_{ij}\}$.

Inference for loglinear models. Estimates and confidence intervals for loglinear model parameters can be obtained using regular GLM methodology, as we discussed last time. Moreover, various tests of different kinds of independence can be framed as either score tests (resulting in Pearson X^2 statistics) or likelihood ratio tests (resulting in G^2 statistics).