MAT 2379, Introduction to Biostatistics

Chapter 14. Categorical Data: Contingency Tables

In Chapter 14, we study data which is classified according to two categorical variables. Suppose that the first variable has r levels and the second variable has c levels. Therefore, there will be $r \times c$ possible ways of classifying the data. The data is displayed into a $r \times c$ *contingency table*.

Example 1. (from p.258 of "Biostatistics. How it works" by Selvin) This is a study of attitudes in rural America. A sample of 130 individuals was collected independently in a small Texas community and asked whether they approve or disapprove of stricter handgun regulations. The same individuals were also classified by two income levels (low and high).

	approve	disapprove	Total
low income	18	48	66 (random)
high income	42	22	64 (random)
Total	60 (random)	70 (random)	130

This example will allow us to illustrate the general theory. In a contingency table we denote by n_{ij} the number of observations that fall into the cell that corresponds to row i and level j. We also calculate the totals for each row and for each column:

$$n_{i\cdot} = n_{i1} + n_{i2} + \ldots n_{ic} \quad \text{total for row } i$$

$$n_{\cdot j} = n_{1j} + n_{2j} + \ldots + n_{rj} \quad \text{total for column } j$$

The total number of observations is $n = n_{1\cdot} + n_{2\cdot} + \ldots + n_{r\cdot} = n_{\cdot 1} + n_{\cdot 2} + \ldots + n_{\cdot c}$.

14.1 Test of Independence

This is the case of when all the row totals and all the column totals are *random*. In this case, we denote by X and Y the two variables under study. Say X has r classes and Y has c classes.

We would like to test the hypothesis H_0 which says that there is no association between the two variables X and Y. In mathematical terms, this can be written as:

$$H_0 : X \text{ and } Y \text{ are independent}$$
$$H_1 : \text{there is an association between } X \text{ and } Y$$

This is called a *test of independence*. To test these hypotheses, we will rephrase H_0 in a more convenient form. Recall from Chapter 5 that two events A and B are *independent* if

$$P(A \cap B) = P(A)P(B)$$

We denote by p_{ij} the probability that a random observation falls into the cell (i, j) of the table (i.e. on row i and column j). We let $p_{i\cdot}$ be the probability that a random observation falls in row i and $p_{\cdot j}$ the probability that a random observation falls in column j. Hypothesis H_0 is restated as:

$$H_0 : p_{ij} = p_{i\cdot}p_{\cdot j} \quad \text{for every } i \text{ and } j$$

Estimators for the probabilities $p_{i\cdot}$ and $p_{\cdot j}$ are:

$$\hat{p}_{i\cdot} = \frac{n_{i\cdot}}{n}, \quad \hat{p}_{\cdot j} = \frac{n_{\cdot j}}{n}$$

If H_0 is true, then $\hat{p}_{ij} = \hat{p}_{i\cdot}\hat{p}_{\cdot j}$ is an estimator for p_{ij}, and the *expected number* of observations that fall into the cell (i, j) is:

$$\hat{E}_{ij} = n\hat{p}_{ij} = n\hat{p}_{i\cdot}\hat{p}_{\cdot j} = n\left(\frac{n_{i\cdot}}{n}\right)\left(\frac{n_{\cdot j}}{n}\right) = \frac{n_{i\cdot}n_{\cdot j}}{n}$$

The expected number $\hat{E}_{ij}$ may be different from the observed number n_{ij} of observations in cell (i, j). If this is the case, then we have some evidence that hypothesis H_0 may not be true. But how different should $\hat{E}_{ij}$ be of n_{ij} in order to be able to reject H_0? To decide if the differences between the observed numbers n_{ij} and the expected numbers $\hat{E}_{ij}$ are significant, we use the test statistic:

$$U_0 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}}$$

It can be proved that U_0 has a *chi-squared distribution* (written χ^2) with $(r-1)(c-1)$ degrees of freedom. The values associated with the χ^2 distribution are given in Table 17.5. The observed value of U_0 in the case of our data is denoted by u_0. The p-value of the test is

$$p-\text{value} = P(U \geq u_0),$$

where U is a random variable with a χ^2 distribution with $(r-1)(c-1)$ d.f. The smaller the p-value, the less likely is becomes that H_0 is true. To find the p-value, we use Table 17.5. If an α-level is specified, we reject H_0 if the p-value is smaller than α.

Example 1. (continued) The null hypothesis H_0 says "there is no association between the opinion about handgun regulations and the income". We suspect that H_0 is not true and we would would like to gain evidence for the alternative H_1 which says that "there is some association between the two variables". We begin by calculated the expected number of observations for each cell.

$$\hat{E}_{11} = \frac{66\cdot 60}{130} = 30.46, \quad \hat{E}_{12} = \frac{66\cdot 70}{130} = 35.34, \quad \hat{E}_{21} = \frac{64\cdot 60}{130} = 29.54, \quad \hat{E}_{22} = \frac{64\cdot 70}{130} = 34.46$$

We put these values in the table underneath the observed values, in parenthesis.

	approve	disapprove	Total
low income	18	48	66
	(30.46)	(35.54)	
high income	42	22	64
	(29.54)	(34.36)	
Total	60	70	130

The observed value of the test statistic is

$$u_0 = \frac{(18-30.46)^2}{30.46} + \frac{(48-35.54)^2}{35.54} + \frac{(42-29.54)^2}{29.54} + \frac{(22-34.36)^2}{34.36} = 19.231$$

The $p-\text{value} = P(U \geq 19.231) < 0.005$ is very small (in Table 17.5, we read $P(U \geq 7.879) = 0.005$).

Since the p-value is so small, we reject H_0. We conclude that there is enough evidence that there in an association between the opinion about the handgun regulations and the income.