Two-Way Tables

Suppose we have a single sample of items, each of which is categorized in two ways. We can arrange the counts of how many times each combination of categories occurs in a *two-way table*.

Here is a table of class (1st, 2nd, 3rd, crew) and sex for people on the Titanic:

ROWS: class COLUMNS: sex

	female	male	ALL
1st	145	180	325
2nd	106	179	285
3rd	196	510	706
crew	23	862	885
ALL	470	1731	2201

Tables of Relative Frequencies

We can divide the counts by the total to get the relative frequencies of items falling in the different cells (in percent here):

ROWS: class COLUMNS: sex

	female	male	ALL
1st	6.59	8.18	14.77
2nd	4.82	8.13	12.95
3rd	8.91	23.17	32.08
crew	1.04	39.16	40.21
ALL	21.35	78.65	100.00

We can view 'class' and 'sex' as random variables, with probabilities given by these relative frequencies.

Testing for Association of Variables

We may be interested in whether the two variables in the table are associated. Viewed in probabilistic terms, we are asking whether the random variables are independent or not.

Recall that if random variables X and Y are independent, then

$$P(X = i \text{ and } Y = j) \;=\; P(X = i)\, P(Y = j)$$

This tells us what the probabilities for an item being in each cell of the table should be, if we know the probabilities for the row and column variables separately.

We will use this relationship to test the null hypothesis of no association.

Expected Cell Counts

Given just the row and column totals, we can fill in the cell counts that we would *expect* (on average), if there is really no association.

If the total count is n, then

$$\begin{aligned} &\text{expected cell count} \\ &\quad = \; n \left(\frac{\text{row total}}{n}\right)\left(\frac{\text{column total}}{n}\right) \\ &\quad = \; \frac{\text{row total} \;\times\; \text{column total}}{n} \end{aligned}$$

We will now compare these expected cell counts to the observed cell counts. The bigger the differences, the more reason we have to doubt the null hypothesis of no association.

The Chi-Squared Test

We will measure how far the observed cell counts differ from what we expect using the following test statistic:

$$\chi^2 = \sum_{\text{all cells}} \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}$$

This statistic has approximately the χ^2 distribution with $(r-1)(c-1)$ degrees of freedom, where r is the number of rows and c is the number of columns. (This is good only if the expected cell counts are fairly large.)

The P-value for our hypothesis test is the probability according to this χ^2 distribution of getting a value for χ^2 at least as great as we actually got.

Titanic Passengers Example

ROWS: class COLUMNS: sex

	female	male	ALL
1st	145	180	325
2nd	106	179	285
3rd	196	510	706
crew	23	862	885
ALL	470	1731	2201

Chi-Square = 349.915, DF = 3, P-Value = 0.000

The P-value is extremely small, as you might guess from just looking at the table.