

Independence

Suppose we observe a number of objects, O_j , $j = 1, \dots, n$ say, each of which is classified with regard to two categorical variables. As an example, suppose the objects are randomly selected persons, and they are classified according to 1) location of residence: *urban, suburban, rural*, 2) education: *no college, four year degree, advanced degree*. Thus for each person we obtain a pair (location, education). Suppose the first categorical variable has r categories and the second has s categories ($r = s = 3$ in the example). When we give labels $j = 1, \dots, r$ and $l = 1, \dots, s$ to the categories, a pair (j, l) may be called a *cell*; we have $r \cdot s$ cells. Each object O_j is classified into one of the cells; as a result we obtain *counts of cell frequencies* M_{jl} , $j = 1, \dots, r$, $l = 1, \dots, s$. The frequencies sum to n : $\sum_{j,l} M_{jl} = n$ since we classified n objects.

The cell counts M_{jl} form a $r \times s$ matrix which is also called a **contingency table**. The contingency table for the location/ education example with $n = 88$ sampled persons looks as follows.

Location	No college	Four-year degree	Advanced degree	Total
Urban	15	12	8	35
Suburban	8	15	9	32
Rural	6	8	7	21
Total	29	35	24	88

In addition to the actual counts M_{jl} , the contingency table also gives the *row totals* $M_{j\bullet}$ in the rightmost column:

$$M_{j\bullet} := \sum_{l=1}^s M_{jl}$$

and the *column totals* in the bottom row:

$$M_{\bullet l} := \sum_{j=1}^r M_{jl}$$

For the *table total* $M_{\bullet\bullet}$ in the bottom right corner we have

$$M_{\bullet\bullet} = \sum_{j,l} M_{jl} = \sum_{j=1}^r M_{j\bullet} = \sum_{l=1}^s M_{\bullet l} = n.$$

We are interested in the question whether there is a connection between the two categorical variables location and education. This is equivalent to the question (as a null hypothesis): for a randomly selected person, are his education and location independent random variables ?

For this one assumes that a randomly selected person or object, when it is classified into one of the cells, generates a bivariate random variable $\mathbf{X} = (X_1, X_2)$ taking values in the finite set of pairs (j, l) , $1 \leq j \leq r$, $1 \leq l \leq s$, where $r, s \geq 2$, with probabilities

$$\Pr((X_1, X_2) = (j, l)) = p_{jl}.$$

These probabilities give the joint distribution of (X_1, X_2) , with marginal distributions

$$P(X_1 = j) = \sum_{l=1}^s p_{jl} =: p_{[1]j}, \quad P(X_2 = l) = \sum_{j=1}^r p_{jl} =: p_{[2]l}.$$

The null hypothesis would be that X_1, X_2 are independent, i.e. the joint distribution is the product of its marginals:

$$p_{jl} = p_{[1]j} \cdot p_{[2]l}, \quad j = 1, \dots, r, \quad l = 1, \dots, s. \quad (101)$$

Thus the n observed objects O_j , $j = 1, \dots, n$ give rise to n i.i.d observed bivariate vectors, $\mathbf{X}_1, \dots, \mathbf{X}_n$, all having the distribution of \mathbf{X} .

We easily recognize this as a hypothesis testing problem about a multinomial distribution. Define an *indicator matrix* \mathbf{Y}_i associated to observation $\mathbf{X}_i = (X_{1i}, X_{2i})$: \mathbf{Y}_i is a $r \times s$ -matrix such that

$$\begin{aligned} \mathbf{Y}_i &= (Y_{i,jl})_{j=1, \dots, r}^{l=1, \dots, s}, \\ Y_{i,jl} &= \begin{cases} 1 & \text{if } (X_{1i}, X_{2i}) = (j, l) \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

Matrices and vectors. The indicator matrices are \mathbf{Y}_i are the analogs of the indicator vectors of dimension k which we defined previously to describe the multinomial distribution. We will identify matrices to vectors by the *stacking operation*: if \mathbf{M} is a $r \times s$ matrix then $\vec{\mathbf{M}}$ is the vector obtained by stacking all the columns of M one below the other; then $\vec{\mathbf{M}}$ has dimension $k = rs$. In what follows we will frequently identify matrices and vectors, but will often not write the stacking operation to simplify notation. Thus we will say that \mathbf{Y}_i has the multinomial distribution

$$\mathbf{Y}_i \sim \mathfrak{M}_{r \times s}(1, \mathbf{p})$$

where \mathbf{p} is the $r \times s$ -matrix of cell probabilities p_{jl} , but 1 is the natural number. This means that for the stacked matrices $\vec{\mathbf{Y}}_i \sim \mathfrak{M}_k(1, \vec{\mathbf{p}})$.

Our total matrices of observed counts is thus

$$\mathbf{M} = \sum_{i=1}^n \mathbf{Y}_i \sim \mathfrak{M}_{r \times s}(n, \mathbf{p})$$

where the multinomial distribution is now immediate (cf. reasoning preceding (91)).

The hypothesis of independence. According to (101), under the independence hypothesis we may write the matrix \mathbf{p} in a special form: define column vectors

$$\mathbf{p}_{[1]} := (p_{[1]j})_{j=1, \dots, r} = \mathbf{p} \mathbf{1}_s, \quad \mathbf{p}_{[2]} := (p_{[2]l})_{l=1, \dots, s} = \mathbf{p}^\top \mathbf{1}_r$$

where $\mathbf{1}_s, \mathbf{1}_r$ are the vectors consisting of 1's of dimension s and r respectively. Thus alternatively we may write the hypothesis of independence

$$\mathbf{p} = \mathbf{p}_{[1]} \mathbf{p}_{[2]}^\top = \mathbf{p} \mathbf{1}_s \mathbf{1}_r^\top \mathbf{p}, \quad (102)$$

or even shorter as

$$\text{rank}(\mathbf{p}) = 1. \quad (103)$$

Indeed any matrix $\mathbf{p}_{(1)}\mathbf{p}_{(2)}^\top$ has rank one. If \mathbf{p} has rank one and is a matrix of probabilities then all rows are multiples of each other which implies a representation $\mathbf{p} = \mathbf{p}_{(1)}\mathbf{p}_{(2)}^\top$. We now have hypotheses

$$H_0 : \mathbf{p} = \mathbf{p}_{[1]}\mathbf{p}_{[2]}^\top$$

$$H_a : \mathbf{p} \neq \mathbf{p}_{[1]}\mathbf{p}_{[2]}^\top$$

The chi square statistic. To measure how much the observed count matrix \mathbf{M} contradicts the null hypothesis, recall the form of the goodness-of-fit chi square statistic in heuristic notation:

$$\chi^2 = \sum_{\text{all cells}} \frac{(\text{observed-expected})^2}{\text{expected}}. \quad (104)$$

Here "expected" means the expected count under the null hypothesis. The observed counts are M_{jl} with expectation under H_0 :

$$EM_{jl} = np_{jl} = np_{[1]j}p_{[2]l}, \text{ or in matrix notation}$$

$$E\mathbf{M} = n\mathbf{p} = n\mathbf{p}_{[1]}\mathbf{p}_{[2]}^\top.$$

Under H_0 we know only the form of \mathbf{p} : $\mathbf{p} = \mathbf{p}_{[1]}\mathbf{p}_{[2]}^\top$ but we do not know the marginal probability vectors $\mathbf{p}_{[1]}$, $\mathbf{p}_{[2]}$. A reasonable procedure is to estimate these under the hypothesis of independence. These estimates can be derived as follows. We would estimate the marginal probability $p_{[1]j}$ of an object falling in to the j -th category (for the first categorical variable) in the usual way, by the sample proportion. The sample proportion in this case is

$$\hat{p}_{[1]j} := n^{-1}M_{j\bullet}$$

since $M_{j\bullet}$ is the count of all objects falling in to the j -th category. Analogously

$$\hat{p}_{[2]l} := n^{-1}M_{\bullet l}$$

This yields an estimated expected value EM_{jl}

$$\widehat{EM}_{jl} = n \cdot \hat{p}_{[1]j} \cdot \hat{p}_{[2]l} = \frac{M_{j\bullet} \cdot M_{\bullet l}}{n}$$

Recalling that the "table total" $M_{\bullet\bullet}$ is equal to n , we may write

$$\widehat{EM}_{jl} = \frac{M_{j\bullet} \cdot M_{\bullet l}}{M_{\bullet\bullet}}.$$

In applied statistics this is usually abbreviated as

$$\text{expected count} = \frac{\text{row total} \cdot \text{column total}}{\text{table total}}$$

but here word "expected count" is misleading since the expression is not an expectation (non-random) but it is random, as a sample based estimate of EM_{jl} under the null hypothesis of independence. The chi square statistic for independence is now

$$\chi^2(\mathbf{M}) = \sum_{\substack{j=1,\dots,r \\ l=1,\dots,s}} \frac{(M_{jl} - n^{-1}M_{j\bullet} \cdot M_{\bullet l})^2}{n^{-1}M_{j\bullet} \cdot M_{\bullet l}}. \quad (105)$$

Its asymptotic null distribution is given by the following.

4.7 Theorem Consider a sequence of random $r \times s$ matrices \mathbf{M}_n having multinomial distribution: $\mathbf{M}_n \sim \mathfrak{M}_{r \times s}(n, \mathbf{p})$ with probability matrix \mathbf{p} satisfying the hypothesis of independence: $\mathbf{p} = \mathbf{p}_{[1]}\mathbf{p}_{[2]}^\top$. Then as $n \rightarrow \infty$ the χ^2 -statistic given by (105) has an asymptotic chi square distribution with $(r-1)(s-1)$ degrees of freedom:

$$\chi^2(\mathbf{M}_n) \rightsquigarrow \chi_{(r-1)(s-1)}^2.$$

This immediately yields an asymptotic α -test for the hypothesis of independence of the random variables X_1, X_2 (each representing a categorical variable), given a matrix of counts \mathbf{M}_n . In the above example, a computation yields

$$\chi^2(\mathbf{M}_n) = 3.01.$$

For $(r-1)(s-1) = 2 \cdot 2 = 4$ degrees of freedom, the table gives the 5% quantile as $\chi_{4,0.05}^2 = 9.488$. Thus the test does not reject, and the data give no statistically significant evidence against independence of location and education.