1 Introduction

The data is Table 1 is a cross-classification of 279 french skiers by experimental group ("placebo" vs. "vitamin C") and occurrence of colds ("yes" and "no"). Both variables have two levels (they are called binary or *dichotomous* variables). For simplicity we denote the two categories of each variable by "1" and "2". The two variables will be denoted by X_1 (experimental group) and X_2 (occurrence of colds).

		Occurrence of cold		
		Yes (1)	No (2)	Row Totals
Treatment	Placebo (1)	31	109	140
	Vitamin C (2)	17	122	139
	Column Totals	48	231	279

Table 1: French skiers data: an example of a 2×2 table.

This table contains four cells corresponding with each combination of categories of the two dichomotous variables:

$$\{(1,1),(1,2),(2,1),(2,2)\}.$$

The cell (i, j) where $1 \leq i \leq 2$ and $1 \leq j \leq 2$ gives the number of skiers n_{ij} that have "Treatment"$= i$ and "Occurrence of cold"$= j$. When we sum over a certain index, we replace by index by "+". In this more general notation, Table 1 can be written as Table 2.

		X_2		
		1	2	Row Totals
X_1	1	n_{11}	n_{12}	n_{1+}
	2	n_{21}	n_{22}	n_{2+}
	Column Totals	n_{+1}	n_{+2}	n_{++}

Table 2: General notation for the counts of a 2×2 table.

The sample size n_{++} is usually called the *grand total* of the table. The row totals give the number of skiers associated with each treatment condition. Similarly, the column totals give the number of skiers who had or did not have a cold. *We will refer to the row and column totals as the one-dimensional marginals*.

2 Two Statistical Models

Since there are two variables, there are only two possibilities:

- The two variables are independent, denoted by $X_1 \perp\!\!\!\perp X_2$.

- The two variables are *not* independent, denoted by $X_1 \not\perp\!\!\!\perp X_2$.

For the French skiers data, $X_1 \perp\!\!\!\perp X_2$ means that taking vitamin C has no effect on whether one gets a cold. The alternative $X_1 \not\perp\!\!\!\perp X_2$ would imply that taking vitamin C will have an effect on the occurrence of colds (although we will not know whether the association is positive or negative). We would like to formally test the hypothesis of independence vs. its alternative. To this end, we need to work with the joint distribution of X_1 and X_2. This joint distribution is specified by the cell probabilities:

$$\begin{aligned} p_{ij} &= \mathsf{P}(X_1 = i, X_2 = j), \\ &= \mathsf{P}(\text{“Treatment”} = i \text{ and “Occurrence of cold”} = j). \end{aligned}$$

We must have:

$$0 < p_{ij} < 1, \text{ for } 1 \le i, j \le 2,$$
$$p_{11} + p_{12} + p_{21} + p_{22} = 1.$$

Therefore, one minus the sum of three cell probabilities will give you the probability of the remaining cell. Look at Table 3. The marginal distribution of X_1 is given by the row totals, while the marginal distribution of X_2 is given by the column totals. To see why this is so, take a look at the following relations:

$$\begin{aligned} \mathsf{P}(X_1 = i) &= \mathsf{P}(X_1 = i, X_2 = 1) + \mathsf{P}(X_1 = i, X_2 = 2), \\ &= p_{i1} + p_{i2}, \\ &= p_{i+}. \end{aligned}$$

We also have:

$$1 = \mathsf{P}(X_1 = 1) + \mathsf{P}(X_1 = 2) = p_{1+} + p_{2+} = p_{++}.$$

Please write the corresponding relations for X_2!

Under the assumption of independence, i.e. $X_1 \perp\!\!\!\perp X_2$, we have:

$$p_{ij} = \mathsf{P}(X_1 = i) \cdot \mathsf{P}(X_2 = j) = p_{i+} \cdot p_{+j}. \tag{1}$$

In other words, each cell probability in Table 3 is given by the product of the corresponding row and column marginal probabilities.

3 Maximum Likelihood Estimates Under Independence

Under independence, the joint distribution of X_1 and X_2 is fully specified by the marginal distributions of X_1 and X_2 — see Eq. (1). These marginal distributions are both binomial:

$$\begin{aligned} X_1 &\sim \mathsf{Bin}(n_{++}; p_{1+}), \\ X_2 &\sim \mathsf{Bin}(n_{++}; p_{+1}). \end{aligned}$$

		X_2		
		1	2	Row Totals
X_1	1	p_{11}	p_{12}	p_{1+}
	2	p_{21}	p_{22}	p_{2+}
	Column Totals	p_{+1}	p_{+2}	$p_{++} = 1$

Table 3: General notation for the cell probabilities of a 2×2 table.

A Binomial distribution $\mathsf{Bin}(m, p)$ is equivalent with a Multinomial distribution $\mathsf{Mult}(m; p, 1-p)$. That is, we have $X_1 \sim \mathsf{Mult}(n_{++}; p_{1+}, 1-p_{1+})$ and $X_2 \sim \mathsf{Mult}(n_{++}; p_{+1}, 1-p_{+1})$. We write:

$$\begin{aligned}
[\mathsf{P}(X_1 = 1)]^x \cdot [\mathsf{P}(X_1 = 2)]^{n_{++}-x} &\propto (p_{1+})^x \cdot (1-p_{1+})^{n_{++}-x}, \text{ for } x = 0, 1, \ldots, n_{++}, \\
[\mathsf{P}(X_2 = 1)]^x \cdot [\mathsf{P}(X_2 = 2)]^{n_{++}-x} &\propto (p_{+1})^x \cdot (1-p_{+1})^{n_{++}-x}, \text{ for } x = 0, 1, \ldots, n_{++}.
\end{aligned}$$

For the counts data from Table 2, the likelihood under independence is given by:

$$[\mathsf{P}(X_1 = 1)]^{n_{1+}} \cdot [\mathsf{P}(X_1 = 2)]^{n_{2+}} \cdot [\mathsf{P}(X_2 = 1)]^{n_{+1}} \cdot [\mathsf{P}(X_2 = 2)]^{n_{+2}} \propto (p_{1+})^{n_{1+}} \cdot (p_{2+})^{n_{2+}} \cdot (p_{+1})^{n_{+1}} \cdot (p_{+2})^{n_{+2}}.$$

Notice that the likelihood depends on the data only through the row and column totals. *This means that the minimal sufficient statistics (MSS, henceforth) of the model of independence are the one-dimensional marginal totals*. By equating to zero the derivatives with respect to p_{1+} and p_{+1}, we obtain the MLEs:

$$\begin{aligned}
\widehat{p}_{1+} &= \frac{n_{1+}}{n_{++}}, \quad \widehat{p}_{2+} = \frac{n_{2+}}{n_{++}}, \\
\widehat{p}_{+1} &= \frac{n_{+1}}{n_{++}}, \quad \widehat{p}_{+2} = \frac{n_{+2}}{n_{++}}.
\end{aligned}$$

It follows that the MLEs for the cell probabilities in Table 3 under the model of independence are:

$$\widehat{p}_{ij} = \widehat{p}_{i+} \cdot \widehat{p}_{+j} = \frac{n_{i+} \cdot n_{+j}}{(n_{++})^2}. \tag{2}$$

Therefore the expected cell counts under the model of independence are:

$$\widehat{m}_{ij} = n_{++} \cdot \widehat{p}_{ij} = \frac{n_{i+} \cdot n_{+j}}{n_{++}}. \tag{3}$$

3.1 Example: Skiers data

Under independence, the expected cells counts are calculated by taking the product of the corresponding row and column totals and dividing it by the grand total. That is, by assuming that vitamin C has no effect

on the occurrence of colds, we should have observed:

$$\begin{aligned}
\widehat{m}_{11} &= \frac{48 \cdot 140}{279} = 24.09, \\
\widehat{m}_{12} &= \frac{231 \cdot 140}{279} = 115.91, \\
\widehat{m}_{21} &= \frac{48 \cdot 139}{279} = 23.91, \\
\widehat{m}_{22} &= \frac{231 \cdot 139}{279} = 115.09.
\end{aligned}$$

4 Maximum Likelihood Estimates Under Interaction

Under the assumption that $X_1 \not\perp\!\!\!\perp X_2$, the joint distribution of X_1 and X_2 is Multinomial:

$$Mult(n_{++}; p_{11}, p_{12}, p_{21}, p_{22}).$$

It follows that the likelihood is proportional with:

$$[\mathsf{P}(X_1 = 1, X_2 = 1)]^{n_{11}} \cdot [\mathsf{P}(X_1 = 1, X_2 = 2)]^{n_{12}} \cdot [\mathsf{P}(X_1 = 2, X_2 = 1)]^{n_{21}} \cdot [\mathsf{P}(X_1 = 2, X_2 = 2)]^{n_{22}},$$

or, equivalently:

$$p_{11}^{n_{11}} \cdot p_{12}^{n_{12}} \cdot p_{21}^{n_{21}} \cdot p_{22}^{n_{22}}.$$

Note that the likelihood depends on the data through all the four cell counts. *This means that the minimal sufficient statistics (MSS) of the model of interaction are the observed cell counts.*.

Remember that we really have only three cells probabilities since they need to add up to one. We equate with zero the derivatives with respect to p_{11}, p_{12} and p_{21}, and obtain the MLEs:

$$\widehat{p}_{ij} = \frac{n_{ij}}{n_{++}}.$$

Therefore, under the model that assumes an interaction between X_1 and X_2, the MLEs of the cell probabilities are obtained by dividing the observed counts by the grand total of the table. It follows that the corresponding expected cell counts are precisely the observed cells counts:

$$\widehat{m}_{ij} = n_{++} \cdot \widehat{p}_{ij} = n_{ij}. \tag{4}$$

For example, under the assumption that vitamin C has a certain effect on the occurrence of colds, the expected cell counts are precisely the counts from Table 1.

5 Testing Independence vs. Interaction

How do we decide whether vitamin C has any effect on colds? In other words, how do we test the null hypothesis

$$H_0 : \quad X_1 \perp\!\!\!\perp X_2,$$

vs. the alternative $X_1 \not\perp\!\!\!\perp X_2$? To answer this question, we can use the likelihood ratio test. That is, we compute the deviance of the model of independence:

$$D(X_1 \perp\!\!\!\perp X_2) = -2\sum_{i=1}^{2}\sum_{j=1}^{2} n_{ij}\log\left(\frac{n_{i+}\cdot n_{+j}}{(n_{++})^2}\right),$$

then the deviance of the interaction model:

$$D(X_1 \not\perp\!\!\!\perp X_2) = -2\sum_{i=1}^{2}\sum_{j=1}^{2} n_{ij}\log\left(\frac{n_{ij}}{n_{++}}\right),$$

The likelihood ratio test statistic is the difference in the deviances of the two models:

$$\begin{aligned} G^2(X_1 \perp\!\!\!\perp X_2 | X_1 \not\perp\!\!\!\perp X_2) &= D(X_1 \perp\!\!\!\perp X_2) - D(X_1 \not\perp\!\!\!\perp X_2), \\ &= -2\sum_{i=1}^{2}\sum_{j=1}^{2} n_{ij}\log\left(\frac{(n_{i+}\cdot n_{+j})/n_{++}}{n_{ij}}\right), \\ &= 2\sum_{i=1}^{2}\sum_{j=1}^{2} n_{ij}\log\left(\frac{n_{ij}}{(n_{i+}\cdot n_{+j})/n_{++}}\right) \end{aligned}$$

This formula is easy to remember in this form:

$$G^2(X_1 \perp\!\!\!\perp X_2 | X_1 \not\perp\!\!\!\perp X_2) = 2\sum_{\text{all cells}} (Observed)\log\left(\frac{Observed}{Expected}\right).$$

For the French skiers data, the likelihood ratio test statistic is:

$$\begin{aligned} G^2(X_1 \perp\!\!\!\perp X_2 | X_1 \not\perp\!\!\!\perp X_2) &= 2\cdot\left(31\log\frac{31}{24.09} + 109\log\frac{109}{115.91} + 17\log\frac{17}{23.91} + 122\log\frac{122}{115.09}\right), \\ &= 4.87 \end{aligned}$$

As you might expect, the asymptotic distribution of $G^2(X_1 \perp\!\!\!\perp X_2 | X_1 \not\perp\!\!\!\perp X_2)$ is Chi-squared with a number of degrees of freedom equal to the difference between the dimensions of the interaction model and the model of independence. We have not see learned the proper log-linear parametrization of these two models yet, but your intuition should tell you that the two models should be different by only one parameter (the same parameter that defines the interaction between X_1 and X_2). Therefore the p-value corresponding with H_0 is given by:

$$\mathsf{P}(\chi_1^2 \geq G^2(X_1 \perp\!\!\!\perp X_2 | X_1 \not\perp\!\!\!\perp X_2)).$$

For the French skiers data, we have $\mathsf{P}(\chi_1^2 \geq 4.87) = 0.027$. Thus we reject independence and conclude that the French skiers data cannot disprove that vitamin C might have some effect on the occurrence of common cold.

Yet another goodness-of-fit test statistic is the X^2 (pronounced X squared):

$$X^2 = \sum_{\text{all cells}} \left(\frac{Observed - Expected}{\sqrt{Expected}}\right)^2.$$

You have already seen this test statistic when we calculated the Pearson residuals. In this context we can assume that each cell count n_{ij} follows a Poisson distribution with mean m_{ij}, i.e.

$$n_{ij} \sim Poisson(m_{ij}).$$

As you might remember for the handout on Poisson regression, the mean of a Poisson is equal with its variance. This implies that the Pearson residual corresponding with the cell (i, j) is obtained by subtracting the mean of n_{ij} and diving by its standard deviation, i.e.

$$\frac{n_{ij} - m_{ij}}{\sqrt{m_{ij}}}.$$

The expression of X^2 represents the sum of the squares of the Pearson residuals corresponding with each cell. In our particular case, the X^2 for testing H_0 is:

$$X^2 = \sum_{i=1}^{2}\sum_{j=1}^{2}\left(\frac{n_{ij} - [(n_{i+} \cdot n_{+j})/n_{++}]}{\sqrt{(n_{i+} \cdot n_{+j})/n_{++}}}\right)^2.$$

The asymptotic distribution of X^2 is coincides with the asymptotic distribution of G^2 (actually, this result holds for any contingency table). Therefore the p-value for testing H_0 based on X^2 is

$$\mathsf{P}(\chi_1^2 \geq X^2).$$

Coming back to the French skiers data, we have

$$X^2 = \left(\frac{31 - 24.09}{\sqrt{24.09}}\right)^2 + \left(\frac{109 - 115.91}{\sqrt{115.91}}\right)^2 + \left(\frac{17 - 23.91}{\sqrt{23.91}}\right)^2 + \left(\frac{122 - 115.09}{\sqrt{115.09}}\right)^2 = 4.806$$

Therefore the corresponding p-value is $\mathsf{P}(\chi_1^2 \geq 4.806) = 0.028$. Remark how similar the values of the two test statistics are.

6 French Skiers Data as a Case-Control Study

So far we have ignored an important aspect of the French skiers data. The number of skiers in the placebo group ($X_1 = 1$) and the number of skiers in the treatment group ($X_1 = 2$) has been fixed in advance. This implies that the marginal distribution of X_1 has not actually been observed, that is, we do not have any information about p_{1+} and p_{2+}. As such, our data consists of the two conditionals:

$$\mathsf{P}(X_2 \mid X_1 = 1) \text{ and } \mathsf{P}(X_2 \mid X_1 = 2).$$

Both conditionals are Binomial, i.e.

$$X_2 \mid X_1 = 1 \sim \mathsf{Bin}(n_{1+}; \mathsf{P}(X_2 = 1 \mid X_1 = 1)), \text{ and } X_2 \mid X_1 = 2 \sim \mathsf{Bin}(n_{2+}; \mathsf{P}(X_2 = 1 \mid X_1 = 2))$$

In our notation, we have

$$\mathsf{P}(X_2 = 1|X_1 = 1) = \frac{\mathsf{P}(X_1 = 1, X_2 = 1)}{\mathsf{P}(X_1 = 1)} = \frac{p_{11}}{p_{1+}}$$

and

$$\mathsf{P}(X_2 = 1|X_1 = 2) = \frac{\mathsf{P}(X_1 = 2, X_2 = 1)}{\mathsf{P}(X_1 = 2)} = \frac{p_{21}}{p_{2+}}$$

If Vitamin C had no effect on the occurrence of colds, the probability of getting a cold in placebo group should be the same as the probability of getting a cold in the treatment group. Therefore we want to test the null hypothesis:

$$H_0 : \quad \mathsf{P}(X_2 = 1|X_1 = 1) = \mathsf{P}(X_2 = 1|X_1 = 2),$$

vs. the alternative $H_A : \quad \mathsf{P}(X_2 = 1|X_1 = 1) \neq \mathsf{P}(X_2 = 1|X_1 = 2)$. The two Binomial conditionals are independent, thus the MLEs of their probabilities of success ($\mathsf{P}(X_2 = 1|X_1 = 1)$ and $\mathsf{P}(X_2 = 1|X_1 = 2)$) are $\frac{n_{11}}{n_{1+}} = \frac{31}{140}$ and $\frac{n_{21}}{n_{2+}} = \frac{17}{139}$, respectively. The MLEs of the probability of success are unbiased, therefore their expected values are precisely the parameters they estimate:

$$\mathsf{E}\left[\frac{n_{11}}{n_{1+}} - \frac{n_{21}}{n_{2+}}\right] = \mathsf{E}\left[\frac{n_{11}}{n_{1+}}\right] - E\left[\frac{n_{21}}{n_{2+}}\right] = \mathsf{P}(X_2 = 1|X_1 = 1) - \mathsf{P}(X_2 = 1|X_1 = 2)$$

They are also independent of each other since the two Binomials are independent. It follows that the variance of their difference is the sum of their individual variances:

$$\begin{aligned}\mathsf{Var}\left[\frac{n_{11}}{n_{1+}} - \frac{n_{21}}{n_{2+}}\right] &= \mathsf{Var}\left[\frac{n_{11}}{n_{1+}}\right] + \mathsf{Var}\left[\frac{n_{21}}{n_{2+}}\right], \\ &= \frac{\mathsf{P}(X_2 = 1|X_1 = 1)(1 - \mathsf{P}(X_2 = 1|X_1 = 1))}{n_{1+}} + \frac{\mathsf{P}(X_2 = 1|X_1 = 2)(1 - \mathsf{P}(X_2 = 1|X_1 = 2))}{n_{2+}}.\end{aligned}$$

If H_0 is true, we would have $\mathsf{P}(X_2 = 1|X_1 = 1) = \mathsf{P}(X_1 = 2|X_2 = 1)$. That is, we could simply combine the data from the Placebo and the Vitamin C groups (it does not make any difference whether you took Vitamin C; the likelihood of getting a cold is the same), i.e.

$$\mathsf{P}(X_2 = 1|X_1 = 1) = \mathsf{P}(X_2 = 1|X_1 = 2) = \mathsf{P}(X_2 = 1).$$

By doing so, our data is actually the row marked "Column Totals" in Table 1. This data corresponds with a Binomial distribution

$$X_2 \sim \mathsf{Bin}(n_{++}; \mathsf{P}(X_2 = 1))$$

The MLE of $\mathsf{P}(X_2 = 1)$ is $\frac{n_{+1}}{n_{++}} = \frac{48}{279}$. It also follows that, under H_0, we have

$$\begin{aligned}\mathsf{E}\left[\frac{n_{11}}{n_{1+}} - \frac{n_{21}}{n_{2+}}\right] &= 0, \\ \mathsf{Var}\left[\frac{n_{11}}{n_{1+}} - \frac{n_{21}}{n_{2+}}\right] &= \mathsf{P}(X_2 = 1)(1 - \mathsf{P}(X_2 = 1))\left(\frac{1}{n_{1+}} + \frac{1}{n_{2+}}\right)\end{aligned}$$

The Central Limit Theorem says that, as the grand total n_{++} goes to infinity, we have

$$\frac{n_{11}}{n_{1+}} - \frac{n_{21}}{n_{2+}} \sim \mathsf{N}\left(0, \mathsf{P}(X_2 = 1)(1 - \mathsf{P}(X_2 = 1))\left(\frac{1}{n_{1+}} + \frac{1}{n_{2+}}\right)\right).$$

Thus an appropriate test statistic for testing H_0 vs H_A is

$$Z = \frac{\frac{n_{11}}{n_{1+}} - \frac{n_{21}}{n_{2+}}}{\sqrt{\frac{n_{+1}}{n_{++}}\left(1 - \frac{n_{+1}}{n_{++}}\right)\left(\frac{1}{n_{1+}} + \frac{1}{n_{2+}}\right)}}.$$

Again, the asymptotic distribution of z under H_0 is $\mathsf{N}(0, 1)$. Thus a p-value for our test is:

$$\mathsf{P}(N(0,1) \geq |Z|) = 2 \cdot \mathsf{P}(N(0,1) \geq |Z|) = 2 \cdot (1 - \Phi(|Z|)).$$

For the French Skiers data, we have $Z = 2.19$ (*please do the calculations yourself to make sure you understand*), thus the p-value is $2 \cdot (1 - \Phi(2.19)) = 0.029$. Remark how similar this p-value is when compared to the p-values we obtained based on the likelihood ratio and X^2 test statistics.

7 Sampling Schemes

Is this similarity of the p-values we computed a pure coincidence? The answer is NO! There is a well-known mathematical result that shows we should actually get the same p-values. This result is related to three sampling schemes used for contingency tables.

A) Poisson Sampling. Each cell count n_{ij} is assumed to follow an independent Poisson distribution with mean m_{ij}. Under this scheme, we do not condition on any quantity.

B) Multinomial Sampling. The cell counts n_{ij} follow a Multinomial distribution $\mathsf{Mult}(n_{++}; (p_{ij})_{i,j})$. Under this scheme, we assume that the grand total n_{++} has been fixed *before* the data was collected. For the French Skiers data, this means we have decided to include 279 skiers in the study.

C) Product-Multinomial Sampling. For each category of a variable, the cell counts are assumed to follow a Multinomial distribution. In the context of the French skiers data, we assume that the row totals $n_{1+} = 140$ and $n_{2+} = 139$ has been fixed *before* the data was collected (actually, this is precisely was happened).Given the row totals we sampled the conditionals

$$X_2 \mid X_1 = 1 \sim \mathsf{Bin}(n_{1+}; \mathsf{P}(X_2 = 1 \mid X_1 = 1)), \text{ and } X_2 \mid X_1 = 2 \sim \mathsf{Bin}(n_{2+}; \mathsf{P}(X_2 = 1 \mid X_1 = 2))$$

If X_2 would have had more than two categories, the Binomial distributions would become Multinomial distributions. The sampling scheme for the entire table is the product of the Multinomial distributions corresponding with each "slice".

The theorem you need to know about states that these three sampling schemes are equivalent. Therefore the expected values (the MLEs) and the goodness-of-fit statistics will be the same no matter what sampling

scheme you assume for your data. Again, in the context of the French skiers data, we *know* that product-multinomial sampling has been used in the collection of the data. However, in your statistical analysis, you can safely assume that the French skiers data has been collected under Poisson or Multinomial sampling. This allows more flexibility in the analysis of your data without raising any questions or doubts related to the validity of the results you report.