CS&SS/STAT/SOC 536: 2×2 Contingency Tables

Adrian Dobra
adobra@uw.edu

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:

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

We must have:

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

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:

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

We also have:

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

Please write the corresponding relations for X_2!

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

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

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

		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.

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:

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

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

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

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

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

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:

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

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:

$$P(\chi^2_1 \geq G^2(X_1 \perp\!\!\!\perp X_2 | X_1 \not\perp\!\!\!\perp X_2)).$$

For the French skiers data, we have $P(\chi^2_1 \geq 4.87) = 0.027$. 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

$$P(\chi^2_1 \geq X^2).$$

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 $P(\chi^2_1 \geq 4.806) = 0.028$. Remark how similar the values of the two test statistics are.

6 The Cross-Product Ratio

Consider the product of the expected values for cells (1,1) and (2,2) divided by the product of the expected values for cell (1,2) and (2,1):

$$\alpha = \frac{m_{11} m_{22}}{m_{21} m_{12}}$$

Since the expected values are always proportional with the corresponding cell probabilities, i.e.

$$m_{ij} = n_{++} \cdot p_{ij},$$

the cross-product can also be expressed as

$$\alpha = \frac{p_{11}p_{22}}{p_{21}p_{12}}$$

The cross-product ratio has a key relevance in the analysis of contingency tables: $\alpha = 1$ under independence $X_1 \perp\!\!\!\perp X_2$. We write:

$$\alpha = \frac{(p_{1+}p_{+1})\cdot(p_{2+}p_{+2})}{(p_{2+}p_{+1})\cdot(p_{1+}p_{+2})} = 1.$$

In order to check if independence could hold for the French skiers data, you calculate the cross-product ratio of the observed counts (remember that these are the expected values under $X_1 \not\perp\!\!\!\perp X_2$):

$$\frac{n_{11}n_{22}}{n_{21}n_{12}} = \frac{31 \cdot 122}{109 \cdot 17} = 2.04$$

Since 2.04 is quite different than 1, independence is unlikely to hold! There are ways to formally test the null hypothesis $H_0 : \alpha = 1$ such as the Fisher's exact test we will talk about later in this handout.

The cross-product ratio is interpreted as an odds ratio. We write:

$$\alpha = \frac{p_{11}/p_{12}}{p_{21}/p_{22}}$$

This represents the odds of being in the first column given that one is in the first row vs. the odds of being in the first column given that one is in the second row. We can also write α as:

$$\alpha = \frac{p_{11}/p_{21}}{p_{12}/p_{22}}.$$

This represents the odds of being in the first row given that one is in the first column vs. the odds of being in the first row given that one is in the second column.

7 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:

$$p(X_2|X_1 = 1) \text{ and } p(X_2|X_1 = 2).$$

Both conditionals are Binomial, i.e.

$$X_2|X_1 = 1 \sim Bin(n_{1+}; p(X_2 = 1|X_1 = 1)), \text{ and } X_2|X_1 = 2 \sim Bin(n_{2+}; p(X_2 = 1|X_1 = 2))$$

In our notation, we have

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

and

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

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:

$$H_0 : \quad p(X_2 = 1|X_1 = 1) = p(X_2 = 1|X_1 = 2),$$

vs. the alternative $H_A : \quad p(X_2 = 1|X_1 = 1) \neq p(X_2 = 1|X_1 = 2)$. The two Binomial conditionals are independent, thus the MLEs of their probabilities of success ($p(X_2 = 1|X_1 = 1)$ and $p(X_2 = 1|X_1 = 2)$) 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:

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

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:

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

If H_0 is true, we would have $p(X_2 = 1|X_1 = 1) = p(X_1 = 2|X_2 = 1)$. 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.

$$p(X_2 = 1|X_1 = 1) = p(X_2 = 1|X_1 = 2) = p(X_2 = 1).$$

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

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

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

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

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

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

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 $N(0, 1)$. Thus a p-value for our test is:

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

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.

8 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 $Mult(n_{++}; (p_{ij})_{i,j})$. 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

$$X_2|X_1 = 1 \sim Bin(n_{1+}; p(X_2 = 1|X_1 = 1)), \text{ and } X_2|X_1 = 2 \sim Bin(n_{2+}; p(X_2 = 1|X_1 = 2))$$

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.

9 Log-linear Models

Let's go back to the formula for the estimated expected cell counts under the model of independence $X_1 \perp\!\!\!\perp X_2$:

$$\widehat{m}_{ij} = \frac{n_{i+} \cdot n_{+j}}{n_{++}}$$

We take the logarithm on both sides to obtain:

$$\log \widehat{m}_{ij} = -\log n_{++} + \log n_{i+} + \log n_{+j}$$

On the righthand side, we see that we have a term corresponding with an overall mean level, a term corresponding with X_1 and a term corresponding with X_2. This is equivalent to having an intercept and two main effects in a generalized linear model. Therefore we consider the following expression for the logarithm of the expected cell values:

$$\log m_{ij} = u + u_{1(i)} + u_{2(j)}, \quad i = 1, 2 \quad j = 1, 2 \tag{5}$$

This expression is a log-linear model because it expresses the logarithm of the expected cell values as a linear function of effects associated with each variable or with their interactions. If you go back to your handout on Poisson regression, you will see the close similarity between Poisson regression and the equivalent form of the log-linear model:

$$m_{ij} = \exp(u + u_{1(i)} + u_{2(j)}), \quad i = 1, 2 \quad j = 1, 2$$

The righthand size involves five parameters: u (grand mean), $u_{1(1)}$ (parameter for category 1 of X_1), $u_{1(2)}$ (parameter for category 2 of X_1), $u_{2(1)}$ (parameter for category 1 of X_2), $u_{2(2)}$ (parameter for category 2 of X_2). In order to make the model identifiable, we must impose two constraints on the u-terms.

One possibility is to consider that category 1 of X_1 and X_2 is baseline and impose:

$$u_{1(1)} = u_{2(1)} = 0.$$

This would imply that the three free parameters of our log-linear model are u, $u_{1(2)}$ and $u_{2(2)}$. While this represents a perfectly valid mathematical choice, it has a serious drawback: the parameter u looses its interpretation as an overall mean level. For this reason, we choose to impose two other constraints:

$$u_{1(1)} + u_{1(2)} = 0, \quad u_{2(1)} + u_{2(2)} = 0 \tag{6}$$

Given this choice, we have:

$$\log m_{11} + \log m_{12} + \log m_{21} + \log m_{22} = 4 \cdot u,$$

which means that u is indeed the overall mean of the logarithm of the expected cell values. Similarly,

$$u + u_{1(i)} = \frac{1}{2}[(u + u_{1(i)} + u_{2(1)}) + (u + u_{1(i)} + u_{2(2)})] = \frac{1}{2}(\log m_{i1} + \log m_{i2})$$

That is, $u + u_{1(i)}$ is the mean of the logarithm of the expected cell counts associated with category i of X_1. The sum $u + u_{2(j)}$ is the mean of the logarithm of the expected cell counts associated with category j of X_2. Moreover, $u_{1(i)}$ and $u_{2(j)}$ represent deviations from the overall mean u.

The log-linear representation of the model of interaction $X_1 \not\perp\!\!\!\perp X_2$ is obtained by adding an interaction term to the log-linear representation of the independence model $X_1 \perp\!\!\!\perp X_2$:

$$\log m_{ij} = u + u_{1(i)} + u_{2(j)} + u_{12(ij)}, \quad i = 1, 2 \quad j = 1, 2$$

In addition to the two constraints on $u_{1(i)}$ and $u_{2(j)}$, we impose four additional constraints on $u_{12(ij)}$:

$$u_{12(1j)} + u_{12(2j)} = 0, \text{ for } j = 1, 2 \quad u_{12(i1)} + u_{12(i2)} = 0, \text{ for } i = 1, 2$$

Please note that every one of these four constraints is a linear combination of the others, so we have added only three constraints. For example, $u_{12(1j)} + u_{12(2j)} = 0$, for $j = 1, 2$ implies that

$$u_{12(11)} + u_{12(21)} + u_{12(12)} + u_{12(22)} = 0.$$

If we also assume $u_{12(11)} + u_{12(12)} = 0$, for the previous relation we obtain $u_{12(21)} + u_{12(22)} = 0$. Therefore, although we added an interaction terms $u_{12(ij)}$ for each cell (i, j), only one of these terms is free. If we know the value of $u_{12(11)}$, we can determine $u_{12(21)} = -u_{12(11)}$, $u_{12(12)} = -u_{12(11)}$ and $u_{12(22)} = u_{12(11)}$. This log-linear model is called *the saturated log-linear model* since it involves the maximum number of interaction terms.

The number of degrees of freedom corresponding with a log-linear model is defined as the number of cells in the table minus the number of free u-terms. The log-linear model of independence has three free u-terms, hence its number of degrees of freedom is $4 - 3 = 1$. The saturated log-linear model has four free u-terms, which implies that it has $4 - 4 = 0$ degrees of freedom.

The null hypothesis $H_0 : X_1 \perp\!\!\!\perp X_2$ can now be written as:

$$H_0 : \quad u_{12(11)} = 0$$

Note that $u_{12(11)} = 0$ implies that $u_{12(21)} = u_{12(12)} = u_{12(22)} = 0$. We also see that the log-linear model of independence and the saturated log-linear model are nested within each other. The test statistics used are G^2 or X^2 – see Section "Testing Independence vs. Interaction". Asymptotically, G^2 or X^2 have a Chi-squared distribution with number of degrees of freedom equal with the difference in the number of degrees of freedom of the two log-linear models being compared. For 2×2 tables, this difference in $1 - 0 = 1$.

10 The Cross-Product Ratio: Revisited

Now we learned that the independence between X_1 and X_2 is expressed as $u_{12(11)} = 0$. Earlier we saw that the independence between X_1 and X_2 is expressed as $\alpha = 1$, where α is the cross-product ratio. It turns out that there is a very simple connection between these two ways to express independence:

$$u_{12(11)} = \frac{1}{4} \log \alpha \tag{7}$$

You see that $\alpha = 1$ is equivalent with $\log \alpha = 0$, which is equivalent with $u_{12(11)} = 0$. Since the cross-product ratio admits an interpretation as odds ratios, it follows that $u_{12(11)}$ shares the same property. Let's quickly prove (7):

$$\begin{aligned}
\frac{1}{4} \log \alpha &= \frac{1}{4} \left(\log m_{11} - \log m_{12} - \log m_{21} + \log m_{22} \right), \\
&= \frac{1}{4} \big(u + u_{1(1)} + u_{2(1)} + u_{12(11)} - u - u_{1(1)} - u_{2(2)} - u_{12(12)} - u - u_{1(2)} - u_{2(1)} - u_{12(21)} \\
&\quad + u + u_{1(2)} + u_{2(2)} + u_{12(22)} \big), \\
&= \frac{1}{4} \cdot 4 \cdot u_{12(11)}, \\
&= u_{12(11)}
\end{aligned}$$

Other u-terms can be expressed in a similar manner, which shows how they should be interpreted:

$$\begin{aligned}
u_{1(1)} &= \frac{1}{4} \log \frac{m_{11} m_{12}}{m_{21} m_{22}}, \\
u_{2(1)} &= \frac{1}{4} \log \frac{m_{11} m_{21}}{m_{12} m_{22}}
\end{aligned}$$

Please write a proof of these two relationships (they are short and easy) and interpret them as odds ratios.

11 Fitting Log-Linear Models in R

We use of the R function "loglin" to fit log-linear models. You need to create an vector in which the counts are ordered such that the index of the first variable varies fastest. We considered that "Treatment" is X_1 and "Occurrence of cold" is X_2. This implies that the counts must be ordered as:

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

The corresponding R commands are:

```
skiers = c(31,17,109,122)
skiers.array = array(skiers,c(2,2))
```

We obtain a 2×2 array that reflects the counts in Table 1 exactly:

```
> skiers.array
     [,1] [,2]
[1,]   31  109
[2,]   17  122
```

We fit the log-linear model of independence of X_1 and X_2:

```
 indep.loglin = loglin(skiers.array,margin=list(1,2),fit=TRUE,param=TRUE)
```

The parameter "fit=TRUE" means that the expected cell values m_{ij} should be calculated. The parameter "param=TRUE" means that the u-terms should be calculated. The parameter "margin=list(1,2)" specifies the minimal sufficient statistics (MSS) of the log-linear model of independence. In this case, the MSS are the row and column totals, that is, the one-dimensional marginals associated with the variables X_1 and X_2. Here is what we obtain:

```
> indep.loglin
$lrt
[1] 4.871697

$pearson
[1] 4.811413

$df
[1] 1
```

This means that (i) the value of the likelihood ratio test statistic G^2 calculated with respect to the saturated log-linear model (i.e., the interaction model) is 4.87; (ii) the value of the X^2 statistic is 4.811 and (iii) the model of independence has one degree of freedom. The rest of the output is interpreted as follows.

```
$margin
$margin[[1]]
[1] 1

$margin[[2]]
[1] 2
```

These are the MSS of the log-linear model that was fit to the data. The first fixed marginal is the one-dimensional marginal corresponding with X_1 (i.e., the row totals). The second fixed marginal is the one-dimensional marginal corresponding with X_2 (i.e., the column totals).

```
$fit
         [,1]     [,2]
[1,] 24.08602 115.9140
[2,] 23.91398 115.0860
```

These are the estimated expected cell values m_{ij} (*please compare with the values we calculated directly with formulas*).

```
$param
$param$`(Intercept)`
[1] 3.963656
```

This is the estimate of the overall mean $u = 3.96$.

```
$param$`1`
[1]  0.003584245 -0.003584245
```

This means that $u_{1(1)} = 0.0036$ and $u_{1(2)} = -0.0036$. Remark that $u_{1(1)} + u_{1(2)} = 0$ according to the constraints we imposed to make the log-linear model identifiable.

```
$param$`2`
[1] -0.7856083  0.7856083
```

This means that $u_{2(1)} = -0.786$ and $u_{2(2)} = 0.786$. Again, the constraint $u_{2(1)} + u_{2(2)} = 0$ is satisfied.

Now let's fit the saturated log-linear model (i.e., the model of interaction $X_1 \not\perp\!\!\!\perp X_2$).

```
saturated.loglin = loglin(skiers.array,margin=list(c(1,2)),fit=TRUE,param=TRUE)
```

For the saturated model, the values of G^2 and X^2 are zero (indicating that the saturated model fits the data perfectly):

```
$lrt
[1] 0

$pearson
[1] 0

$df
[1] 0
```

We also see that the saturated model does not have any degrees of freedom (again, we fitted four expected cell values with four parameters).

```
$margin
$margin[[1]]
[1] 1 2
```

This means that the MSS of the saturated log-linear model is the marginal associated with the variables X_1 and X_2. Since we only have two variables, this is the original French skiers table.

```
$fit
     [,1] [,2]
[1,]   31  109
[2,]   17  122
```

The estimated expected cell values under the saturated log-linear model are the counts in the original table.

```
$param
$param$`(Intercept)`
[1] 3.940642

$param$`1`
[1]  0.1220252 -0.1220252

$param$`2`
[1] -0.8070421  0.8070421

$param$`1.2`
           [,1]       [,2]
[1,]  0.1783618 -0.1783618
[2,] -0.1783618  0.1783618
```

These are the estimates of the u-terms. We have: $u = 3.94$, $u_{1(1)} = 0.122$, $u_{1(2)} = -0.122$, $u_{2(1)} = -0.807$, $u_{2(2)} = 0.807$, $u_{12(11)} = 0.178$, $u_{12(12)} = -0.178$, $u_{12(21)} = -0.178$ and $u_{12(22)} = 0.178$.

To obtain the p-value for testing $H_0 : \quad u_{12(11)} = 0$ using the likelihood ratio test statistic G^2, we use

```
1-pchisq(indep.loglin$lrt-saturated.loglin$lrt,
               indep.loglin$df-saturated.loglin$df)
```

We obtain 0.0273. To perform the same test based on the X^2 test statistic, we use

```
1-pchisq(indep.loglin$pearson-saturated.loglin$pearson,
               indep.loglin$df-saturated.loglin$df)
```

We obtain 0.0283. *You should run all these R commands on your own to make sure you understand them!*

12 Fitting Log-Linear Models in R with "GLM"

There is another method to fit log-linear models in R by employing our old friend from regression, the function "glm". To this end, we have to create a file in which the French skiers data are represented as follows:

```
 y Treatment Cold
31 1 1
109 1 2
17 2 1
122 2 2
```

This is exactly the format of the data we used when fitting regression models of various kinds. Each row is associated with a cell in the table. The count that appears in each cell is given in the first column of the corresponding row. Let's read in the data and fit the log-linear model of independence:

```
skiers = read.table('skiers.txt',header=TRUE);
skiers.indep = glm(y ~ factor(Treatment) + factor(Cold),family=poisson,
                                data=skiers)
> skiers.indep

Call:  glm(formula = y ~ factor(Treatment) + factor(Cold), family = poisson,
                     data = skiers)

Coefficients:
       (Intercept)  factor(Treatment)2       factor(Cold)2
          3.181632           -0.007168            1.571217

Degrees of Freedom: 3 Total (i.e. Null);  1 Residual
Null Deviance:     135.5
Residual Deviance: 4.872  AIC: 34
```

We specify "family=poisson" because we can assume the table has been observed under Poisson sampling. You have just fitted the log-linear model of independence (5) under the constraints:

$$u_{1(1)} = u_{2(1)} = 0.$$

When you used the function "loglin", you fitted model (5) under the sum constraints (6). The output above says that

$$u = 3.181, \quad u_{1(2)} = -0.007, \quad u_{2(2)} = 1.571$$

Category "1" associated with "Treatment" and "Occurrence of colds" has been considered to be baseline, hence the corresponding u-terms are fixed at zero (i.e. $u_{1(1)} = u_{2(1)} = 0$). You can convince yourselves you have fitted the same model of independence by looking at the expected cell counts you obtain:

```
 > fitted(skiers.indep)
        1         2         3         4
 24.08602 115.91398  23.91398 115.08602
```

The order of the expected cell counts are precisely the same order you specified the cells in the data file:

$$\widehat{m}_{11} = 24.086, \quad \widehat{m}_{12} = 115.91, \quad \widehat{m}_{21} = 23.91, \quad \widehat{m}_{22} = 115.08$$

These are precisely the same expected values we found before. You can fit the log-linear interaction model (i.e., the saturated model) by including an interaction term between "Treatment" and "Occurrence of colds":

```
 skiers.saturated = glm(y ~ factor(Treatment) + factor(Cold) +
                                        factor(Treatment):factor(Cold),
                                        family=poisson,
                                       data=skiers)
> skiers.saturated

Call:  glm(formula = y ~ factor(Treatment) + factor(Cold) +
                            factor(Treatment):factor(Cold),
                        family = poisson, data = skiers)

Coefficients:
                      (Intercept)                 factor(Treatment)2
                           3.4340                            -0.6008
                    factor(Cold)2  factor(Treatment)2:factor(Cold)2
                           1.2574                             0.7134

Degrees of Freedom: 3 Total (i.e. Null);  0 Residual
Null Deviance:     135.5
Residual Deviance: -5.773e-15  AIC: 31.13
```

Remark that there is only one non-zero parameter describing the interaction between "Treatment" and "Occurrence of colds". Again, category "1" associated with "Treatment" and "Occurrence of colds" has been considered to be baseline, hence the following u-terms associated with category "1" were fixed at zero to make the model identifiable:

$$u_{1(1)} = u_{2(1)} = u_{12(11)} = u_{12(12)} = u_{12(21)} = 0$$

The non-zero u-terms from the output above are

$$u = 2.72, \quad u_{1(2)} = -0.6, \quad u_{2(2)} = 1.257, \quad u_{12(22)} = 0.713$$

We look at the estimated expected cell values

```
 > fitted(skiers.saturated)
  1   2   3   4
 31 109  17 122
```

and see that they coincide with the observed cell counts. Therefore we fit the saturated log-linear model by imposing a different set of zero constraints for the u-terms. Numerically, both sets of constraints are equivalent, thus we can get information about the a log-linear model by fitting them using "glm" AND "loglin".

13 The Connection Between Log-linear and Logit Models

The log-linear models of independence and interaction translate into regression models. For example, let's determine the logit of "Treatment" (X_1) given "Occurrence of colds" (X_2) from the

log-linear model of independence (5):

$$\begin{aligned}
\log \frac{P(X_1 = 2|X_2 = j)}{P(X_1 = 1|X_2 = j)} &= \log \frac{P(X_1 = 2, X_2 = j)}{P(X_1 = 1, X_2 = j)}, \\
&= \log \frac{p_{2j}}{p_{1j}}, \\
&= \log \frac{m_{2j}}{m_{1j}}, \\
&= \log m_{2j} - \log m_{1j}, \\
&= (u + u_{1(2)} + u_{2(j)}) - (u + u_{1(1)} + u_{2(j)}), \\
&= u_{1(2)} - u_{1(1)}
\end{aligned}$$

This model is the intercept-only regression:

$$\log \frac{P(X_1 = 2|X_2)}{P(X_1 = 1|X_2)} = \beta_0 \tag{8}$$

Remark that the logit of "Treatment" does not dependent on "Occurrence of colds" since the two variables are assumed independent. In particular, we obtain:

$$\log \frac{P(X_1 = 2|X_2 = 1)}{P(X_1 = 1|X_2 = 1)} = \log \frac{P(X_1 = 2|X_2 = 2)}{P(X_1 = 1|X_2 = 2)} = u_{1(2)} - u_{1(1)} = -0.007 \tag{9}$$

Pay attention to the estimates of the u-terms obtained using the function "loglin" and "glm". From "loglin" we obtained $u_{1(1)} = 0.0036$ and $u_{1(2)} = -0.0036$, thus $u_{1(2)} - u_{1(1)} = -2 \cdot 0.0036 = -0.007$. From "glm" we obtained $u_{1(1)} = 0$ and $u_{1(2)} = -0.007$, thus, once again, $u_{1(2)} - u_{1(1)} = -0.007$.

Now we fit model (8) as a logistic regression. First we transform the counts in the French skiers data in records, then we fit the intercept-only logistic regression with "Treatment" as the response:

```
mydata = matrix(c(rep(c(1,1),31),rep(c(1,2),109),
                          rep(c(2,1),17),rep(c(2,2),122)),
                          ncol=2,byrow=TRUE)
mylogit = glm(factor(mydata[,1])~1,family=binomial(link=logit))
> mylogit

Call:  glm(formula = factor(mydata[, 1]) ~ 1, family = binomial(link = logit))

Coefficients:
(Intercept)
  -0.007168

Degrees of Freedom: 278 Total (i.e. Null);  278 Residual
Null Deviance:     386.8
Residual Deviance: 386.8  AIC: 388.8
```

The output reveals that the estimate of β_0 in model (8) is -0.007, exactly as before.

Next let's determine the same logit from the saturated log-linear model:

$$\begin{aligned}\log\frac{P(X_1=2|X_2=j)}{P(X_1=1|X_2=j)} &= \log m_{2j} - \log m_{1j},\\ &= (u + u_{1(2)} + u_{2(j)} + u_{12(2j)}) - (u + u_{1(1)} + u_{2(j)} + u_{12(1j)}),\\ &= (u_{1(2)} - u_{1(1)}) + (u_{12(2j)} - u_{12(1j)})\end{aligned}$$

The logit of "Treatment" depends on the "Occurrence of colds" since the two variables are assumed to interact with each other. We obtain:

$$\log\frac{P(X_1=2|X_2=1)}{P(X_1=1|X_2=1)} = (u_{1(2)} - u_{1(1)}) + (u_{12(21)} - u_{12(11)}) \tag{10}$$

The output of "loglin" shows that: $u_{1(1)} = 0.122$, $u_{1(2)} = -0.122$, $u_{12(11)} = 0.178$, $u_{12(21)} = -0.178$, hence

$$(u_{1(2)} - u_{1(1)}) + (u_{12(2j)} - u_{12(1j)}) = -0.122 - 0.122 - 0.178 - 0.178 = -0.6$$

On the other hand, the output of "glm" shows that $u_{1(1)} = u_{12(11)} = u_{12(21)} = 0$ and $u_{1(2)} = -0.6$, which leads to the same logit value as before. Similarly, we have:

$$\log\frac{P(X_1=2|X_2=2)}{P(X_1=1|X_2=2)} = (u_{1(2)} - u_{1(1)}) + (u_{12(22)} - u_{12(12)}) \tag{11}$$

The output of "loglin" shows that $u_{12(12)} = -0.178$ and $u_{12(22)} = 0.178$, thus

$$(u_{1(2)} - u_{1(1)}) + (u_{12(22)} - u_{12(12)}) = -0.122 - 0.122 + 0.178 + 0.178 = 0.112$$

The output of "glm" shows that $u_{1(1)} = 0$, $u_{1(2)} = -0.6$, $u_{12(12)} = 0$ and $u_{12(22)} = 0.713$, thus we obtain the same value as before:

$$(u_{1(2)} - u_{1(1)}) + (u_{12(22)} - u_{12(12)}) = -0.6 + 0.713 = 0.112$$

Remark that equations (10) and (11) are precisely the logistic regression model containing an intercept and a slope for X_2:

$$\log\frac{P(X_1=2|X_2)}{P(X_1=1|X_2)} = \beta_0 + \beta_1 X_2 \tag{12}$$

Let's fit this logistic regression model directly:

```
 > mylogit = glm(factor(mydata[,1])~factor(mydata[,2]),
                            family=binomial(link=logit))
> mylogit

Call:  glm(formula = factor(mydata[, 1]) ~ factor(mydata[, 2]),
                 family = binomial(link = logit))
```

```
Coefficients:
         (Intercept)  factor(mydata[, 2])2
             -0.6008                0.7134

Degrees of Freedom: 278 Total (i.e. Null);  277 Residual
Null Deviance:     386.8
Residual Deviance: 381.9  AIC: 385.9
```

The output indicates that $\beta_0 = -0.6, \beta_1 \times (X_2 = 1) = 0$ and $\beta_1 \times (X_2 = 2) = 0.713$.

In the previous lectures we have discussed variable selection in regression. When we have determined we prefer the saturated model $X_1 \not\perp\!\!\!\perp X_2$ to the independence model $X_1 \perp\!\!\!\perp X_2$, we have actually performed variable selection for the regression of X_1 given X_2. That is $X_1 \perp\!\!\!\perp X_2$ corresponds to the null logistic regression of X_1 (i.e., the regression that contains only the intercept, see equation (8)), while $X_1 \not\perp\!\!\!\perp X_2$ corresponds to the logistic regression of X_1 on X_2 (i.e., the regression that contains an intercept and a slope parameter for X_2, see equation (12)). This means that our preferred logistic regression for X_1 is given by (12) and not by (8).

14 Exact Testing

There is another method to determine how well the independence model $X_1 \perp\!\!\!\perp X_2$ fits the data (remember that the saturated model $X_1 \not\perp\!\!\!\perp X_2$ *always* fits the data perfectly). This method is called exact testing and involves conditioning on the row and column totals. As we have already explained, the row and column totals are the minimal sufficient statistics of the independence model. Assume you are given these row and column totals, but you do NOT know the values of the four counts in Table 1. More precisely, assume you are given Table 4 instead of Table 1.

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

Table 4: French skiers data with missing cell counts.

What values can each of the four cell counts take given the row and column totals? The set of possible values of each cell is characterized by their Frechet lower and upper bounds:

$$\max\{0, n_{i+} + n_{+j} - n_{++}\} \leq n_{ij} \leq \min\{n_{i+}, n_{+j}\}.$$

For the French skiers data, we obtain:

$$\begin{aligned}
\max\{0, 48 + 140 - 279\} = 0 \leq n_{11} \leq 48 = \min\{48, 140\},\\
\max\{0, 231 + 140 - 279\} = 92 \leq n_{12} \leq 140 = \min\{231, 140\},\\
\max\{0, 48 + 139 - 279\} = 0 \leq n_{21} \leq 48 = \min\{48, 139\},\\
\max\{0, 231 + 139 - 279\} = 91 \leq n_{22} \leq 139 = \min\{231, 139\}
\end{aligned}$$

We are actually interested in the set of 2×2 tables with the same row and column totals as the French skiers data:

$$T = \{n' = (n'_{ij})_{1 \leq i,j \leq 2} : n'_{1+} = 140, n'_{2+} = 139, n'_{+1} = 48, n'_{+2} = 231\}.$$

Each counts in each table in T must belong to the Frecher lower and upper bounds we determined before. In particular, we see that each cell can take the same number of values (i.e., the number of integer values between each lower and upper bound):

$$49 = 48 - 0 + 1 = 140 - 92 + 1 = 48 - 0 + 1 = 138 - 91 + 1$$

This suggests that the set T contains precisely 49 tables. Let's assume that the count in the (1,1) cell takes a value x, where x is an integer between 0 and 48. It follows that the count in the (1,2) cell must be $140 - x$ and the cell in the (2,1) cell must be $48 - x$. Thus the count in the $(2, 2)$ cell must be $139 - (48 - x) = 91 + x$. Therefore the set T is described as:

$$T = \{(x, 140 - x, 48 - x, 91 + x) : x \in \{0, 1, \ldots, 48\}\}.$$

The French skiers data corresponds with $x = 31$. Under Multinomial sampling, the probability of each table $n' \in T$ is

$$P(n') = \frac{n'_{++}!}{n'_{11}!n'_{12}!n'_{21}!n'_{22}!} p_{11}^{n'_{11}} p_{12}^{n'_{12}} p_{21}^{n'_{21}} p_{22}^{n'_{22}}$$

Under the model of independence, the cell probabilities as written as a function of the u-terms:

$$\log p_{ij} = u + u_{1(i)} + u_{2(j)}$$

It follows that:

$$\begin{aligned}
\prod_{i,j} p_{ij}^{n'_{ij}} &= \prod_{i,j} \exp\left(n'_{ij}u + n'_{ij}u_{1(i)} + n'_{ij}u_{2(j)}\right),\\
&= \exp\left(n'_{++}u + n'_{1+}u_{1(1)} + n'_{2+}u_{1(2)} + n'_{+1}u_{2(1)} + n'_{+2}u_{2(2)}\right).
\end{aligned}$$

But this last expression is constant for all the tables in T since all the tables in T have the same grand total and the same row and column marginal totals. Therefore:

$$\prod_{i,j} p_{ij}^{n'_{ij}} = \textit{constant}, \text{ for all } n' \in T. \tag{13}$$

We consider the probability of a table $n' \in T$ *conditional on the set* T:

$$\begin{aligned} P(n'|T) &= \frac{P(n')}{\sum_{n'' \in T} P(n'')}, \\ &= \frac{\frac{n'_{++}!}{n'_{11}!n'_{12}!n'_{21}!n'_{22}!} p_{11}^{n'_{11}} p_{12}^{n'_{12}} p_{21}^{n'_{21}} p_{22}^{n'_{22}}}{\sum_{n'' \in T} \frac{n''_{++}!}{n''_{11}!n''_{12}!n''_{21}!n''_{22}!} p_{11}^{n''_{11}} p_{12}^{n''_{12}} p_{21}^{n''_{21}} p_{22}^{n''_{22}}}, \\ &= \frac{\frac{1}{n'_{11}!n'_{12}!n'_{21}!n'_{22}!}}{\sum_{n'' \in T} \frac{1}{n''_{11}!n''_{12}!n''_{21}!n''_{22}!}} \end{aligned}$$

The above equation represents the hypergeometric distribution, although not in a form you might have seen before. The cell probabilities cancel because of equation (13). This fact will become extremely important when the tables are sparse (i.e., contain many zeros). In these situations, the asymptotic distributions of G^2 or X^2 might not be Chi-squared. These are the situations when you will have to perform exact testing instead of making use of the asymptotic tests we developed so far.

Let's denote by $n \in T$ the observed table (in our case the French skiers data). The exact p-value for testing the model of independence is given by the sum of the hypergeometric probabilities of the tables that have a G^2 or an X^2 test statistic greater or equal to the G^2 or X^2 test statistic corresponding with the observed table n:

$$\begin{aligned} \text{Exact p-value for } G^2 &= \sum_{\{n' \in T : G^2(n') \geq G^2(n)\}} P(n'|T), \\ \text{Exact p-value for } X^2 &= \sum_{\{n' \in T : X^2(n') \geq X^2(n)\}} P(n'|T) \end{aligned}$$

For the French skiers data we would have to calculate G^2 and X^2 for all 49 tables in T, calculate the hypergeometric probabilities of each table, then figure out the sum of the hypergeometric probabilities of the tables that give higher values of G^2 or X^2 than the corresponding values of n.

REMEMBER: We have defined exact p-values for 2×2 *tables, but similar definitions hold for any contingency table and any log-linear model. In most cases the exact enumeration of* T *is not possible and Markov chain Monte Carlo methods would have to be employed.*

15 Exact Computation of Exact P-values

The French skiers data is small enough to allow us to generate all the tables with the same row and column totals, calculate their hypergeometric probabilities, then calculate exactly the exact p-value based on G^2 and X^2. This computation is performed in the R code that follows. You can use the same code to calculate exact p-values for any 2×2 table by changing the Frechet lower and upper bounds corresponding with cell (1,1), the number of tables consistent with these bounds

(this number is one plus the difference between the upper and the lower bounds for cell (1,1)) and the formula that identifies a table based on the count of the (1,1) cell and the row/column totals.

```
#change this function to reflect your 2x2 table
probtable <- function(x)
{
  return(lfactorial(x)+lfactorial(140-x)+lfactorial(48-x)+lfactorial(91+x))
}

#this is the number of 2x2 tables
ntables = 49
LowerBoundCell11 = 0
UpperBoundCell11 = 48
ObservedCountCell11 = 31

#calculate the probabilities of all tables
alltablesprob = numeric(ntables)
for(x in LowerBoundCell11:UpperBoundCell11)
{
  alltablesprob[x+1] <- -probtable(x)
}

alltablesprob <- exp(alltablesprob-max(alltablesprob))
x <- alltablesprob/sum(alltablesprob)
alltablesprob <- x

#see how the normalized probabilities look like
barplot(alltablesprob,names=0:(ntables-1),
               xlab='Count in cell (1,1)',
               ylab='Hypergeometric probability of table')

#calculate G^2 for all tables
G2 = numeric(ntables)
for(x in LowerBoundCell11:UpperBoundCell11)
{
G2[x-LowerBoundCell11+1] = loglin(array(c(x,140-x,48-x,91+x),c(2,2)),
                                                    margin=list(1,2))$lrt
}

#calculate X^2 for all tables
X2 = numeric(ntables)
for(x in LowerBoundCell11:UpperBoundCell11)
{
```

```
X2[x-LowerBoundCell11+1] = loglin(array(c(x,140-x,48-x,91+x),c(2,2)),
                                                    margin=list(1,2))$pearson
}

#calculate the exact p-value based on G^2
ExactPvalueG2 = 0
for(x in LowerBoundCell11:UpperBoundCell11)
{
if(G2[x-LowerBoundCell11+1]>=G2[ObservedCountCell11-LowerBoundCell11+1])
{
ExactPvalueG2 = ExactPvalueG2 + alltablesprob[x-LowerBoundCell11+1]
}
}

#calculate the exact p-value based on X^2
ExactPvalueX2 = 0
for(x in LowerBoundCell11:UpperBoundCell11)
{
if(X2[x-LowerBoundCell11+1]>=X2[ObservedCountCell11-LowerBoundCell11+1])
{
ExactPvalueX2 = ExactPvalueX2 + alltablesprob[x-LowerBoundCell11+1]
}
}

> ExactPvalueX2
[1] 0.03849249
> ExactPvalueG2
[1] 0.03849249
```

We obtained the exact p-values based on G^2 and X^2 to be equal with 0.038. Remember that the corresponding asymptotic p-values were 0.027 and 0.028, respectively. We can still reject the model of independence based on the exact p-value. The p-values (of any kind, asymptotic or exact) associated with G^2 and X^2 *happen* to be the same because the values of these test statistics *happen* to be the same for the French skiers data. In general, G^2 and X^2 could be quite different, which leads to quite different asymptotic or exact p-values.

For higher-dimensional tables it is rarely possible to perform an exact calculation of the exact p-values by generating *all* tables consistent with a set of fixed marginals. As such, the exact p-values need to be estimated and the corresponding computations can be done with the "exactLoglinTest" R package written by Brian Caffo. You need to represent your contingency table in the format you used when fitting log-linear models using "glm". We will learn how to use of this R package in the next handouts.

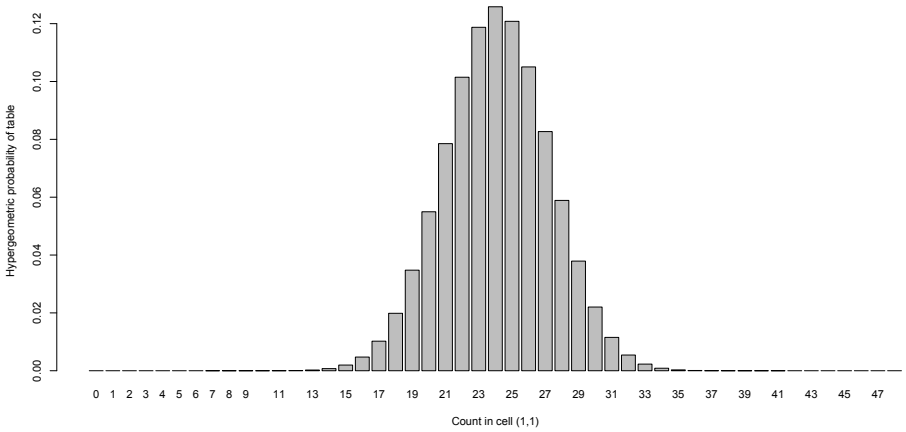

Figure 1: Hypergeometric probabilities of each 2×2 table having the same row and column totals as the French skiers data. The observed French skiers table is obtained for by setting the count in cell (1,1) to $x = 31$.

16 Fisher's Exact Test

The computation of the exact p-value for 2×2 tables is performed using the R function "fisher.test". To use this function you need to represent your table as an array (the same array "skiers.array" needed to use the function "loglin"). We would want to test whether there is an association between Vitamin C and the occurrence of colds vs. the alternative of no association using the cross-product ratio α. That is, we want to test

$$H_0 : \alpha = 1 \quad vs. \quad H_A : \alpha \neq 1$$

We make the following call in R:

```
> fisher.test(skiers.array, simulate.p.value=TRUE, B=1e5)

Fisher's Exact Test for Count Data

data:  skiers.array
p-value = 0.03849
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 1.026674 4.154449
sample estimates:
odds ratio
  2.035861
```

The output shows that the odds ratio α is 2.04 for the French skiers data (the same value we calculated directly). We also see that the estimated exact p-value is 0.038, hence we reject H_0

(*Remark that this is the same value we previously obtained when computing exact p-values based on X^2 or G^2. This is simply a coincidence; in general you will not get the same estimates when computing exact p-values associated with different test statistics. Moreover, the exact p-value for a test statistics is not necessarily equal with the asymptotic p-value for the same test statistic. They could be close, but should NOT necessarily be the same*). This implies that there is an association between Vitamin C and the occurrence of colds.

Fisher's exact test allows us to go one step further and actually test whether this association is positive or negative. In other words, we would expect Vitamin C to *decrease* the occurrence of colds, not only to *influence* it (i.e., decrease OR increase it). Therefore, we want to see whether the odds of getting a cold in the placebo group (i.e., p_{11}/p_{12}) is *larger* than the odds of getting a cold in the Vitamin C group (i.e., p_{21}/p_{22}). Equivalently, this means we would like to test:

$$H_0 : \alpha = 1 \quad vs. \quad H_A : \alpha > 1$$

We make the following call in R:

```
> fisher.test(skiers.array, simulate.p.value=TRUE, alternative = 'greater',B=1e5)

Fisher's Exact Test for Count Data

data:  skiers.array
p-value = 0.02052
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval:
 1.134558      Inf
sample estimates:
odds ratio
  2.035861
```

The output shows that the exact p-value for this test is 0.02, hence we can reject H_0 and decide that the data provides evidence that Vitamin C *decreases* the effect of colds. It is a useful exercise to calculate the exact p-value for testing

$$H_0 : \alpha = 1 \quad vs. \quad H_A : \alpha < 1$$

We make the call in R:

```
> fisher.test(skiers.array, simulate.p.value=TRUE, alternative = 'less',B=1e5)

Fisher's Exact Test for Count Data

data:  skiers.array
p-value = 0.991
alternative hypothesis: true odds ratio is less than 1
95 percent confidence interval:
```

```
 0.000000 3.723449
sample estimates:
odds ratio
  2.035861
```

This time the exact p-value is estimated to be 1, therefore we fail to reject H_0 given the alternative that Vitamin C *increases* the occurrence of colds.

17 Bayesian Testing

We denote by $\mathcal{D}$ the observed cell counts $\{n_{ij} : 1 \leq i \leq 2, 1 \leq j \leq 2\}$. The data $\mathcal{D}$ could have arisen under two hypotheses $H_1 : X_1 \perp\!\!\!\perp X_2$ (independence) or $H_2 : X_1 \not\perp\!\!\!\perp X_2$ (interaction). Before seeing the data we do not have any opinion about which of these two hypotheses is more probable. As such, we assume that H_1 and H_2 are a priori equally likely:

$$pr(H_1) = pr(H_2) = 0.5$$

The Bayes' theorem shows that the ratio of the posterior probabilities of H_1 and H_2 is

$$\frac{pr(H_2|\mathcal{D})}{pr(H_1|\mathcal{D})} = B_{21}\frac{pr(H_2)}{pr(H_1)},$$

where B_{21} is the Bayes factor

$$B_{21} = \frac{pr(\mathcal{D}|H_2)}{pr(\mathcal{D}|H_1)}.$$

Therefore the Bayes factor represents the ratio of posterior odds of H_1 to its prior odds. Since we assume that the prior odds of H_1 are equal to 1, B_{12} is equal with the posterior odds. The densities $pr(\mathcal{D}|H_1)$ and $pr(\mathcal{D}|H_1)$ are called marginal or integrated likelihoods and are obtain by averaging the likelihood over all possible values of the parameters under H_1 and H_2. Remember that likelihood ratio tests are based on maximizing the likelihood over the possible values of the parameters under each hypothesis. We average with respect to prior distributions for the model parameters. How do we obtain suitable priors? To answer this question, consider the fictive data in Table 5:

		X_2		
		1	2	Row Totals
X_1	1	α	α	2α
	2	α	α	2α
	Column Totals	2α	2α	4α

Table 5: Fictive 2×2 data.

Table 5 has all its four cell entries equal to α indicating that, *before seeing any data*, we do not have any knowledge about which combinations of categories are more or less likely.

Under the hypothesis of interaction H_2, the Multinomial likelihood is proportional with:

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

The joint prior distribution for cell probabilities is assumed to be Dirichlet$(\alpha, \alpha, \alpha, \alpha)$:

$$pr(p_{11}, p_{12}, p_{21}, p_{22}|\alpha) = \frac{\Gamma(4\alpha)}{\Gamma(\alpha)^4} p_{11}^{\alpha-1} p_{12}^{\alpha-1} p_{21}^{\alpha-1} p_{22}^{\alpha-1}.$$

The Dirichlet distribution is conjugate to the Multinomial likelihood, hence the joint posterior distribution of cell probabilities is again Dirichlet$(n_{11} + \alpha, n_{12} + \alpha, n_{21} + \alpha, n_{22} + \alpha)$:

$$pr(p_{11}, p_{12}, p_{21}, p_{22}|\mathcal{D}, \alpha) = \frac{\Gamma(n_{++} + 4\alpha)}{\Gamma(n_{11} + \alpha)\Gamma(n_{12} + \alpha)\Gamma(n_{21} + \alpha)\Gamma(n_{22} + \alpha)} p_{11}^{n_{11}+\alpha-1} p_{12}^{n_{12}+\alpha-1} p_{21}^{n_{21}+\alpha-1} p_{22}^{n_{22}+\alpha-1}.$$

Your intuition should be that the posterior distribution of cell probabilities is obtained by augmenting the observed cell counts in Table 2 with the fictive prior cell values in Table 5. The end result is a 2×2 table with grand total $n_{++} + 4\alpha$ and cell entries $n_{ij} + \alpha$. It is recommended that the prior has a small "weight" in the posterior distribution than the data. To this end, we take $\alpha = 0.25$ which leads to a prior table with a grand total of 1 and each cell entry equal to 0.25. These values are much smaller than the observed counts in Table 1.

The marginal likelihood $pr(\mathcal{D}|H_2)$ is obtained by integrating out the cell probabilities from the Multinomial likelihood:

$$pr(\mathcal{D}|H_2) = \int_{\{(p_{11},p_{12},p_{21},p_{22}):p_{11}+p_{12}+p_{21}+p_{22}=1, p_{ij}>0\}} p_{11}^{n_{11}} \cdot p_{12}^{n_{12}} \cdot p_{21}^{n_{21}} \cdot p_{22}^{n_{22}} pr(p_{11}, p_{12}, p_{21}, p_{22}|\alpha) dp_{11} dp_{12} dp_{21} dp_{22}.$$

It is easy to see that $pr(\mathcal{D}|H_2)$ is equal with the ratio of the normalizing constants of the Dirichlet prior and posterior distributions:

$$pr(\mathcal{D}|H_2) = \frac{\Gamma(n_{11} + \alpha)\Gamma(n_{12} + \alpha)\Gamma(n_{21} + \alpha)\Gamma(n_{22} + \alpha)}{\Gamma(n_{++} + 4\alpha)} \frac{\Gamma(4\alpha)}{\Gamma(\alpha)^4}.$$

Under the hypothesis of independence H_1, the likelihood is proportional with:

$$(p_{1+})^{n_{1+}} \cdot (p_{2+})^{n_{2+}} \cdot (p_{+1})^{n_{+1}} \cdot (p_{+2})^{n_{+2}}.$$

We assume independent Dirichlet$(2\alpha, 2\alpha)$ priors for the row and column marginal probabilities:

$$\begin{aligned} pr(p_{1+}, p_{2+}) &= \frac{\Gamma(4\alpha)}{\Gamma(2\alpha)^2} p_{1+}^{2\alpha-1} p_{2+}^{2\alpha-1}, \\ pr(p_{+1}, p_{+2}) &= \frac{\Gamma(4\alpha)}{\Gamma(2\alpha)^2} p_{+1}^{2\alpha-1} p_{+2}^{2\alpha-1}. \end{aligned}$$

The joint posterior distribution of the row and column cell probabilities is the product of Dirichlet$(n_{1+} + 2\alpha, n_{2+} + 2\alpha)$ and Dirichlet$(n_{+1} + 2\alpha, n_{+2} + 2\alpha)$:

$$\begin{aligned} pr(p_{1+}, p_{2+}, p_{+1}, p_{+2}|\mathcal{D}, \alpha) &= \frac{\Gamma(n_{++} + 4\alpha)^2}{\Gamma(n_{1+} + 2\alpha)\Gamma(n_{2+} + 2\alpha)\Gamma(n_{+1} + 2\alpha)\Gamma(n_{+2} + 2\alpha)} \times \\ &\times (p_{1+})^{n_{1+}+2\alpha-1} \cdot (p_{2+})^{n_{2+}+2\alpha-1} \cdot (p_{+1})^{n_{+1}+2\alpha-1} \cdot (p_{+2})^{n_{+2}+2\alpha-1}. \end{aligned}$$

Once again, the marginal likelihood associated with H_1 is equal with the ratio of posterior and prior Dirichlet normalizing constants:

$$pr(\mathcal{D}|H_1) = \frac{\Gamma(n_{1+} + 2\alpha)\Gamma(n_{2+} + 2\alpha)\Gamma(n_{+1} + 2\alpha)\Gamma(n_{+2} + 2\alpha)}{\Gamma(n_{++} + 4\alpha)^2} \frac{\Gamma(4\alpha)^2}{\Gamma(2\alpha)^4}.$$

Kass and Raftery (1995) suggest the following guidelines related to the evidence provided by the data in favor of H_2 and against H_1 – see Table 6.

Table 6: Interpretation of the evidence provided by Bayes factors.

$2 \log_e B_{21}$	Evidence against H_1 AND in favor of H_2
0 to 2	Not worth more than a bare mention
2 to 6	Positive
6 to 10	Strong
>10	Very strong

The Bayes factor B_{21} is calculated with the following R code for the French skiers data:

```
bayes.test = function(var1,var2)
{
#two-way table
obstable = table(var1,var2);
#first one-way marginal
rowtable = table(var1);
#second one-way marginal
columntable = table(var2);

#calculate the log-marginal likelihood under the saturated log-linear model
alpha = 0.25;
logmargSaturated = lgamma(4*alpha)-4*lgamma(alpha);
logmargSaturated = logmargSaturated +sum(lgamma(as.vector(obstable+alpha)));
logmargSaturated = logmargSaturated -lgamma(sum(as.vector(obstable+alpha)));

#calculate the log-marginal likelihood under the log-linear model of independence
logmargIndep = 2*lgamma(4*alpha)-4*lgamma(2*alpha);
logmargIndep = logmargIndep + sum(lgamma(as.vector(rowtable+2*alpha)));
logmargIndep = logmargIndep+sum(lgamma(as.vector(columntable+2*alpha)));
logmargIndep = logmargIndep - lgamma(sum(as.vector(rowtable+2*alpha)));
logmargIndep = logmargIndep -lgamma(sum(as.vector(columntable+2*alpha)));
return(2*(logmargSaturated-logmargIndep));
}

mydata = matrix(c(rep(c(1,1),31),rep(c(1,2),109),
  rep(c(2,1),17),rep(c(2,2),122)),
```

```
ncol=2,byrow=TRUE);

> bayes.test(mydata[,1],mydata[,2])
[1] -1.264203
```

Therefore $2\log_e B_{21} = -1.264$ which is equivalent to $2\log_e B_{12} = 2\log_e \frac{pr(\mathcal{D}|H_1)}{pr(\mathcal{D}|H_2)} = 1.264$. From Table 6 we see that the French skiers data provides weak evidence in favor of H_1 (independence) and against H_2 (interaction). This seems to go in the opposite direction of the frequentist conclusions we have reached before.