GENERALISED LINEAR MODELLING
LECTURE NOTES: ANALYSIS OF CONTINGENCY TABLES

I. Introduction

In the case of logistic regression, we took a categorical response variable and "transformed" it into a numerical response by using proportions. This was made possible by the fact that the real response variable we were dealing with had only two categories. For instance, in our initial anaesthetic depth example (Section 2, Example 1), we grouped together individuals with equivalent predictor values (dosages) and then summarised our response as the proportion within each group who responded to stimulus. Similarly, for the premature baby exercise (Tutorial Week 9, Exercise 1) we simply used each individual baby and summarised the response again as a proportion, however, this time that proportion was either zero or one. In either case, we could make this summarisation in a straightforward manner and without any loss of information since there were only two categories, and a single proportion sufficed to describe the data in each "group". We now wish to expand our examination of categorical response variables to the case where there are more than two possible outcomes.

Example 1 - Premarital Sex and Upbringing: A study on the sexual values of first-year university students in America was conducted in the late 1960's. The objective of the study was to relate sexual values to various socio-economic and geographical covariates. A sample of 1407 students was taken and (among many other variables) the geographical region where the student was raised was recorded, along with their view on the acceptability of premarital sex. Both of the variables here are clearly categorical, and the raw data would consist of 1407 pairs, one for each sampled student, indicating in which region the corresponding individual was raised and what their view was on the acceptability of premarital sex. The most convenient way to summarise this sort of data is in the form of *counts* collected in a *contingency table*. The actual observed data were:

View on Premarital Sex	Region Raised: East Coast	South	Midwest	West Coast	Outside U.S.	Total
Very Acceptable	82	201	30	169	13	495
Somewhat Acceptable	72	251	20	96	7	446
Somewhat Unacceptable	49	122	26	36	4	237
Very Unacceptable	36	149	13	28	3	229
Total	239	723	89	329	27	1407

Clearly, the question of interest is whether the breakdown of student views on premarital sex is the same regardless of the region in which the students were raised. Unlike the case of a categorical response with only two levels, we cannot reduce the response to a single proportion now without loss of information, and if we summarise using the three proportions it would take to maintain all the information (note that we would not need the proportion in each of the four categories, since the last proportion is constrained by the values of the other three to ensure that all the proportions together sum to unity, which is why we could get away with only one proportion in the case of a categorical variable with two levels), we have a response variable which is a vector. So, we have to think of a somewhat different way to deal with analysing this data.

Typically, two-way contingency table data (i.e., count data broken down into the "cells" created by the cross-classification of two categorical variables) can be broadly classified according to two important criteria. First, we can make the distinction between "nominal" categorical variables and "ordinal" categorical variables. Nominal variables are ones for which the ordering of the groups is uninformative, such as is the case for the region in which an individual was raised. Ordinal

variables, on the other hand, are ones for which the ordering is meaningful, for example, the level of acceptability of premarital sex in the preceding example. Of course, we can always simply ignore the ordering structure of an ordinal variable and treat it as a nominal one, and this is the focus we shall adopt for the first part of our discussion. The more advanced procedures which are designed to take into account the extra information inherent in the ordering structure of ordinal variables will be briefly discussed later in this chapter.

We can also make a distinction based on how the data were gathered. Typically, there will be two basic ways in which we can obtain two-way contingency table data. First, we can simply sample an overall total number of subjects, which means that the overall total sample size, n, is the only thing fixed by our sampling design. Alternatively, we can sample a preset number of individuals from each category of one of our variables. This means that one set of the marginal totals (which by convention is usually set to be the column totals) is fixed by our sample design. We shall discuss each of these situations, but first we need to set up some nomenclature and define appropriate probability structures.

II. Probability Structure

For two categorical variables A and B with R and C levels, respectively, we will denote the observed data in an $R \times C$ contingency table as:

Level of A	Level of B: 1	2	$\cdots$	C	Total
1	Y_{11}	Y_{12}	$\cdots$	Y_{1C}	$Y_{1\bullet}$
2	Y_{21}	Y_{22}	$\cdots$	Y_{2C}	$Y_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
R	Y_{R1}	Y_{R2}	$\cdots$	Y_{RC}	$Y_{R\bullet}$
Total	$Y_{\bullet 1}$	$Y_{\bullet 2}$	$\cdots$	$Y_{\bullet C}$	$n = Y_{\bullet\bullet}$

where Y_{ij} is the *observed count* for the $(i,j)^{\text{th}}$ *cell* of the contingency table, while $Y_{\bullet j}$ and $Y_{i\bullet}$ are the *marginal* j^{th} column and i^{th} row totals, respectively. For completeness, we note that we can accomodate count data which tabulates responses for more than two categorical variables into *multi-way* contingency tables, and we denote their cell counts using the obvious extension of the above notation; namely, by using additional subscripts on the Y's, one for each categorical variable, and again a "$\bullet$" indicates that summation across the corresponding index has been performed.

The most straightforward probability model for $R \times C$ contingency table data is under the scheme where only the overall total, n, is fixed. In this situation, the most general model is based on the *multinomial* distribution, which defines the chance that each sampled subject contributes to the count in the $(i,j)^{\text{th}}$ cell as π_{ij}. We can then write the probability mass function (which is also the likelihood function) for the observed counts $Y = (Y_{11}, Y_{12}, \ldots, Y_{RC})$ as:

$$f(Y|n) = n! \prod_{i=1}^{R} \prod_{j=1}^{C} \frac{\pi_{ij}^{Y_{ij}}}{Y_{ij}!},$$

where we require that the probability values π_{ij} satisfy $0 \leq \pi_{ij} \leq 1$ and $\sum_{i=1}^{R} \sum_{j=1}^{C} \pi_{ij} = 1$. We will also denote the probability that an individual count contributes to the i^{th} row or the j^{th} column as $\pi_{i\bullet} = \sum_{j=1}^{C} \pi_{ij}$ and $\pi_{\bullet j} = \sum_{i=1}^{R} \pi_{ij}$, respectively. Note that the requirement that the π_{ij}'s sum to unity implies that the $\pi_{i\bullet}$'s and $\pi_{\bullet j}$'s, the *marginal probabilities* must also sum to unity.

In the case where the column totals, $Y_{\bullet j}$, are fixed by the sampling design, we can think of the observations in each column as coming from its own multinomial distribution, and thus the overall

likelihood of the observed counts is based on the *product multinomial* distribution:

$$f(Y|Y_{\bullet 1},\ldots,Y_{\bullet C}) = \prod_{j=1}^{C} Y_{\bullet j}! \prod_{i=1}^{R} \frac{\pi_{ij}^{Y_{ij}}}{Y_{ij}!},$$

where we now require that $0 \leq \pi_{ij} \leq 1$ and $\sum_{i=1}^{R} \pi_{ij} = \pi_{\bullet j} = 1$ for every value of the second index, i.e., the probabilities in each column must sum to unity. Note that we write the values that are assumed to be fixed by the sampling scheme after the vertical bar in the notation to remind ourselves which design was used to gather the data. As a final point, we note that the above distributions arise quite naturally out of conditional probability calculations for the Poisson distribution; however, a demonstration of this concept is outside the scope of our current course. Nonetheless, it adds another reason why the logarithmic link function which we will employ in the next section is useful, since this is exactly the canonical link for the Poisson distribution.

III. Log-Linear Models for Two-way Contingency Tables

We shall start with the case in which the data has been gathered with only the total n being fixed. In such a situation, it can be shown (and is also intuitively clear) that:

$$E(Y_{ij}) = n\pi_{ij}.$$

Typically, any hypothesis of interest regarding the parameters π_{ij} in this situation are in the form of multiplicative models for the expected cell frequencies. For instance, the most common hypothesis in this situation is that the two categorical variables have no association with each other, implying that the probability of a sampled individual contributing to a particular row would be independent of their chance of contributing to a particular column. In other words, the hypothesis of interest is to test whether $\pi_{ij} = \pi_{i\bullet}\pi_{\bullet j}$ or not. Clearly then, the model that we are interested in examining is:

$$E(Y_{ij}) = n\pi_{i\bullet}\pi_{\bullet j} \quad \Longrightarrow \quad \begin{aligned} \log\{E(Y_{ij})\} &= \log n + \log \pi_{i\bullet} + \log \pi_{\bullet j} \\ &= \log n + \tau_i + \alpha_j, \end{aligned}$$

where we have defined new parameters $\tau_i = \log \pi_{i\bullet}$ and $\alpha_j = \log \pi_{\bullet j}$ and the fact that there are constraints on the $\pi_{i\bullet}$'s and the $\pi_{\bullet j}$'s (i.e., that they must sum to unity) translates into the necessity for constraints on the new parameters. The model structure here is clearly reminiscent of the normal theory ANOVA, and thus the same possible constraint systems are equally applicable. Now, the general model can also be written in "ANOVA-type" form as:

$$\begin{aligned} \log\{E(Y_{ij})\} = \log(n\pi_{ij}) = \log\left(\frac{n\pi_{i\bullet}\pi_{\bullet j}\pi_{ij}}{\pi_{i\bullet}\pi_{\bullet j}}\right) &= \log n + \log \pi_{i\bullet} + \log \pi_{\bullet j} + \log\left(\frac{\pi_{ij}}{\pi_{i\bullet}\pi_{\bullet j}}\right) \\ &= \log n + \tau_i + \alpha_j + \gamma_{ij}. \end{aligned}$$

where we have defined $\gamma_{ij} = \log(\pi_{ij}/\pi_{i\bullet}\pi_{\bullet j})$ and we again must translate the constraints on the π_{ij}'s into constraints on the new parameters, and we shall again use the parallel structure of the interactive ANOVA model as our guide.

So, given this structure, we see that the hypothesis of independence between the two variables under study can be framed as $H_0 : \gamma_{ij} = 0$ for all $1 \leq i \leq R$ and $1 \leq j \leq C$ [NOTE: Recall that there are implied constraints on the γ_{ij}'s and thus this hypothesis does not actually concern RC "effective" parameters; see the reparameterisation below]. Further, we note that if we construct

appropriate indicator variables as predictors, we can write our model as a *log-linear* one, since:

$$\begin{aligned}\log\{E(Y_{ij})\} &= \log n + \tau_i + \alpha_j + \gamma_{ij} \\ &= \beta_0 + \beta_1 z_{1,ij} + \ldots + \beta_{R-1} z_{R-1,ij} + \beta_R w_{1,ij} + \ldots + \beta_{R+C-2} w_{C-1,ij} \\ &\quad + \beta_{1,1} z_{1,ij} w_{1,ij} + \ldots + \beta_{R-1,C-1} z_{R-1,ij} w_{C-1,ij} \\ &= \beta_0 + Z_{ij}\beta_{(1)} + W_{ij}\beta_{(2)} + (Z \otimes W)_{ij}\beta_{(3)},\end{aligned}$$

where the β's are defined in terms of $\log n$ and the τ_i's, α_j's and γ_{ij}'s as:

$$\begin{aligned}\beta_0 &= \log n + \tau_1 + \alpha_1 + \gamma_{11} \\ \beta_i &= (\tau_{i+1} - \tau_1) + (\gamma_{i+1,1} - \gamma_{11}) \qquad 1 \le i \le R-1 \\ \beta_{R+j} &= (\alpha_{j+2} - \alpha_1) + (\gamma_{1,j+2} - \gamma_{11}) \qquad 0 \le j \le C-2 \\ \beta_{i,j} &= \gamma_{i+1,j+1} - \gamma_{i+1,1} - \gamma_{1,j+1} + \gamma_{11} \qquad 1 \le i \le R-1;\ 1 \le j \le C-1,\end{aligned}$$

and the z's and w's are the appropriate indicators of the row or column, respectively, to which the observed count Y_{ij} belongs (e.g., $z_{k,ij}$ is the indicator for the $(k+1)^{\text{st}}$ row and $w_{k,ij}$ is the indicator for the $(k+1)^{\text{st}}$ column). Moreover, the vectors Z_{ij}, W_{ij} and $(Z \otimes W)_{ij}$ are defined in analogy to the case for normal two-way ANOVA, so that $Z_{ij} = (z_{1,ij}, \ldots, z_{R-1,ij})$, $W_{ij} = (w_{1,ij}, \ldots, w_{C-1,ij})$,

$$\begin{aligned}(Z \otimes W)_{ij} = (&z_{1,ij}w_{1,ij}, z_{2,ij}w_{1,ij}, \ldots, z_{R-1,ij}w_{1,ij}, z_{1,ij}w_{2,ij}, \ldots, z_{R-1,ij}w_{2,ij}, \\ &\ldots, z_{1,ij}w_{C-1,ij}, \ldots, z_{R-1,ij}w_{C-1,ij}),\end{aligned}$$

and the parameter vectors are similarly defined as $\beta_{(1)} = (\beta_1, \ldots, \beta_{R-1})^T$, $\beta_{(2)} = (\beta_R, \ldots, \beta_{R+C-2})^T$ and $\beta_{(3)} = (\beta_{1,1}, \ldots, \beta_{R-1,C-1})^T$. In this final form, it can be shown (employing the implied constraints on the γ_{ij}'s which derive from the constraints on the π_{ij}'s) that our null hypothesis of independence becomes $H_0 : \beta_{(3)} = 0$, which shows explicitly that the hypothesis of interest concerns only $(R-1)(C-1)$ parameters. Note that there are exactly RC parameters in the full model, and only RC observed counts as well. [TECHNICAL NOTE: Actually, there are only $RC-1$ "effective" data points since the counts are constrained to sum up to the fixed total n, however, the fact that the fixed and known value $\log n$ appears in the definition of the β's means that there are only effectively $RC-1$ of them as well.] Fortunately, the multinomial distribution is an exponential family with dispersion parameter $\phi = 1$, and thus extra degrees of freedom are not required for its estimation, just as was the case for binomial and Poisson GLMs. Unfortunately, the `glm()` function of *S-Plus* does not support a `family=multinomial` option, and so we will have to devise another way to estimate the β's and test our hypothesis of independence. We shall do this in the following section.

Before we do, however, we will complete this section with a parallel discussion for the case in which the sampling design has fixed each of the column totals, $Y_{\bullet j}$. For such situations, it turns out (and again is intuitively clear) that,

$$E(Y_{ij}) = Y_{\bullet j}\pi_{ij}.$$

In this situation, the question of interest is generally one of *homogeneity* of the columns. In other words, we would like to know if the distribution of the counts among the levels of the row variable is the same within each of the columns. If the columns are homogeneous, then for each $i \in \{1, \ldots, R\}$ we must have $\pi_{i1} = \pi_{i2} = \cdots = \pi_{iC} = \pi_{i\bullet}$, since the average value of a collection of equal numbers

is just equal to the common value of each member of the collection. So, the model for homogeneity is:

$$E(Y_{ij}) = Y_{\bullet j}\pi_{i\bullet} \qquad \Longrightarrow \qquad \log\{E(Y_{ij})\} = \log(Y_{\bullet j}\pi_{i\bullet}) = \log\left(\frac{nY_{\bullet j}\pi_{i\bullet}}{n}\right)$$
$$= \log n + \log \pi_{i\bullet} + \log\left(\frac{Y_{\bullet j}}{n}\right) = \log n + \tau_i + \alpha_j,$$

which is exactly of the form for the previous case. Note that the definition of the α parameters is somewhat different, however, we are going to be much more interested in hypothesis testing than parameter estimation for contingency table data. Finally, we note that the full model can be written in the same form that it was for the previous case, since we have:

$$\begin{aligned}\log\{E(Y_{ij})\} &= \log(Y_{\bullet j}\pi_{ij}) \\ &= \log\left(\frac{nY_{\bullet j}\pi_{i\bullet}\pi_{ij}}{n\pi_{i\bullet}}\right) \\ &= \log n + \log \pi_{i\bullet} + \log\left(\frac{Y_{\bullet j}}{n}\right) + \log\left(\frac{\pi_{ij}}{\pi_{i\bullet}}\right) \\ &= \log n + \tau_i + \alpha_j + \gamma_{ij} \\ &= \beta_0 + Z_{ij}\beta_{(1)} + W_{ij}\beta_{(2)} + (Z \otimes W)_{ij}\beta_{(3)}\end{aligned}$$

And thus, the test for homogeneity of the column distributions has the same null hypothesis as the test for independence did; namely, $H_0 : \gamma_{ij} = 0$ or equivalently $H_0 : \beta_{(3)} = 0$. Note, however, that there are some distinctions between the two models when we want to test other hypotheses. For example, suppose we wanted to see if there was a "column effect", that is, whether the distribution in the population of interest was evenly (or *uniformly*) distributed among the column variable categories (i.e., we want to test whether the α's should be retained in the model or not). In the case of sampling with only a fixed total, we can do this in the obvious way. However, for the case of sampling with fixed column margins, it makes no sense to test whether the observed column proportions are uniform or not, since we have set these totals during the sampling process. In other words, the α_j's have "fixed" values under this sampling scheme and their values cannot be of any use in making inferences about the population under study. In such a situation, the α_j's are referred to as *constraint parameters* and they must be retained in any model we fit to the observed data.

IV. Maximum Likelihood and Tests of Independence and Homogeneity

We start with the case of sampling with only a fixed total. In this situation, we can use the multinomial probability function defined above and the fact that $\mu_{ij} = E(Y_{ij}) = n\pi_{ij}$ to show that the log-likelihood function for $\mu = (\mu_{11}, \ldots, \mu_{RC})$ is given by:

$$\begin{aligned}l(\mu|n) &= \log n! + \sum_{i=1}^{R}\sum_{j=1}^{C}\left\{Y_{ij}\log\pi_{ij} - \log Y_{ij}!\right\} \\ &= \log n! + \sum_{i=1}^{R}\sum_{j=1}^{C}\left\{Y_{ij}\log\left(\frac{\mu_{ij}}{n}\right) - \log Y_{ij}!\right\} \\ &= \sum_{i=1}^{R}\sum_{j=1}^{C} Y_{ij}\log\mu_{ij} + \log n! - \sum_{i=1}^{R}\sum_{j=1}^{C}\left\{Y_{ij}\log n + \log Y_{ij}!\right\} \\ &= \sum_{i=1}^{R}\sum_{j=1}^{C} Y_{ij}\log\mu_{ij} + d(Y),\end{aligned}$$

where the function $d(Y)$ is appropriately defined. We note that, as we shall see, the form of the function $d(Y)$ is unimportant in future calculations because it does not have any reference to the parameters, π_{ij} (in just the same way that the $d(Y_i, \phi)$ function in the definition of the form of the likelihood for an exponential family with dispersion had no reference to the β's, and thus its form was unimportant in tests and confidence statements regarding these parameters.)

Using this log-likelihood it is not difficult to show, and indeed it is intuitively clear, that the maximum likelihood estimates of the π_{ij}'s under the full or *saturated* model (i.e., the model in which the only constraint on the π_{ij}'s is that they sum to unity) are just

$$\hat{\pi}_{ij} = \frac{Y_{ij}}{n} \qquad \Longrightarrow \qquad \hat{\mu}_{ij} = Y_{ij}.$$

From this result, it again follows quite intuitively that the maximum likelihood estimates for the row and column probabilities, $\pi_{i\bullet}$ and $\pi_{\bullet j}$, are just:

$$\hat{\pi}_{i\bullet} = \sum_{j=1}^{C} \hat{\pi}_{ij} = \frac{1}{n} \sum_{j=1}^{C} Y_{ij} = \frac{Y_{i\bullet}}{n}; \qquad \hat{\pi}_{\bullet j} = \sum_{i=1}^{R} \hat{\pi}_{ij} = \frac{1}{n} \sum_{i=1}^{R} Y_{ij} = \frac{Y_{\bullet j}}{n}.$$

Now, this last result shows that under the model of independence, where we assume that $\pi_{ij} = \pi_{i\bullet}\pi_{\bullet j}$, the fitted values for the expectations μ_{ij} are then:

$$\tilde{\mu}_{ij} = n\hat{\pi}_{i\bullet}\hat{\pi}_{\bullet j} = \frac{Y_{i\bullet}Y_{\bullet j}}{n},$$

where we use the "˜" to repesent fitted values under the model of independence. Another common notation for these fitted values under the assumption of independence is $E_{ij} = \tilde{\mu}_{ij}$, where the "$E$" stands for *expected*, indicating that this is the value which would be expected under the independence assumption. When this notation is employed, it is common to replace Y_{ij} by O_{ij} where the "O" stands for *observed*, the value being the observed count in the $(i,j)^{\text{th}}$ cell of the contingency table.

From this, we see that the deviance statistic for the test of independence (recalling that $\phi = 1$ for multinomial distributions) is just:

$$\begin{aligned} D^\star(\tilde{\mu}, \hat{\mu}) &= \frac{D(\tilde{\mu}, Y) - D(\hat{\mu}, Y)}{1} \\ &= 2\{l(\hat{\mu}|n) - l(\tilde{\mu}|n)\} \\ &= 2\left\{ \sum_{i=1}^{R}\sum_{j=1}^{C} Y_{ij} \log \hat{\mu}_{ij} + d(Y) - \sum_{i=1}^{R}\sum_{j=1}^{C} Y_{ij} \log \tilde{\mu}_{ij} - d(Y) \right\} \\ &= 2 \sum_{i=1}^{R}\sum_{j=1}^{C} Y_{ij} \log\left(\frac{\hat{\mu}_{ij}}{\tilde{\mu}_{ij}}\right) \\ &= 2 \sum_{i=1}^{R}\sum_{j=1}^{C} O_{ij} \log\left(\frac{O_{ij}}{E_{ij}}\right). \end{aligned}$$

$\sim (1-\alpha)\ \chi^2$

So, we can test the null hypotheses of independence (i.e., test whether the "smaller" model is adequately describing the observed count data) at significance level α by comparing this deviance statistic to the $(1-\alpha)$-quantile of an appropriate chi-squared distribution. As usual, the appropriate chi-squared distribution has degrees of freedom equal to the difference in the number of parameters in the small and large models. Clearly, for the saturated model there are RC parameters, while for

the independence model there are $1 + (R-1) + (C-1) = R + C - 1$. [TECHNICAL NOTE: As pointed out earlier, there are actually only $RC - 1$ "effective" parameters in the saturated model, however, there are also only $R + C - 2$ "effective" parameters in the independence model, again due to the presence of the fixed value $\log n$ in the definition of the β's.] Therefore, the appropriate degrees of freedom for our hypothesis test is $RC - R - C + 1 = (R-1)(C-1)$, so that we will reject the null hypothesis of independence if

$$D^\star(\tilde{\mu}, \hat{\mu}) \geq \chi^2_{(R-1)(C-1)}(1-\alpha).$$

It turns out that there is an approximately equivalent test statistic which is more commonly used in practice. To derive it, we must first note a useful fact about logarithms. If δ is a number close to zero, then

$$\log(1+\delta) \approx \delta - \tfrac{1}{2}\delta^2.$$

A demonstration of this fact is not difficult, but requires some cleverness, and we present it here only for the sake of completeness. We start by defining the function $h(\delta) = \delta^{-1}\log(1+\delta)$, and then note that the linearisation technique shows that, since $\delta \approx 0$:

$$h(\delta) \approx h(0) + h'(0)(\delta - 0).$$

Now, implicitly differentiating the equation $\delta h(\delta) = \log(1+\delta)$ with respect to δ twice shows that:

$$\begin{aligned}
h(\delta) + \delta h'(\delta) = (1+\delta)^{-1} \quad &\Longrightarrow \quad h(0) + 0h'(0) = (1+0)^{-1} \\
&\Longrightarrow \quad h(0) = 1 \\
h'(\delta) + \{h'(\delta) + \delta h''(\delta)\} = -(1+\delta)^{-2} \quad &\Longrightarrow \quad h'(0) + h'(0) + 0h''(0) = -(1+0)^{-2} \\
&\Longrightarrow \quad h'(0) = -\tfrac{1}{2}.
\end{aligned}$$

So, using these values in the linearisation yields:

$$\delta^{-1}\log(1+\delta) = h(\delta) \approx 1 - \tfrac{1}{2}\delta \qquad \Longrightarrow \qquad \log(1+\delta) \approx \delta - \tfrac{1}{2}\delta^2.$$

Now, if our independence hypothesis is true then the E_{ij} values should be close to the observed counts, O_{ij}, so that:

$$\frac{O_{ij}}{E_{ij}} = 1 + \delta_{ij} \qquad \Longrightarrow \qquad \delta_{ij} = \frac{O_{ij} - E_{ij}}{E_{ij}},$$

where δ_{ij} is a number close to zero (and indeed becomes smaller and smaller as n, the sample size, gets larger and larger). In fact, it can be shown that δ_{ij} is close enough to zero that δ_{ij}^3 is negligible in size. So, we can now write the deviance statistic as:

$$\begin{aligned}
D^\star(\tilde{\mu}, \hat{\mu}) &= 2\sum_{i=1}^{R}\sum_{j=1}^{C} O_{ij}\log\left(\frac{O_{ij}}{E_{ij}}\right) = 2\sum_{i=1}^{R}\sum_{j=1}^{C} E_{ij}(1+\delta_{ij})\log(1+\delta_{ij}) \\
&\approx 2\sum_{i=1}^{R}\sum_{j=1}^{C} E_{ij}(1+\delta_{ij})\left(\delta_{ij} - \tfrac{1}{2}\delta_{ij}^2\right) = \sum_{i=1}^{R}\sum_{j=1}^{C} E_{ij}(2\delta_{ij} + \delta_{ij}^2 - \delta_{ij}^3) \\
&\approx \sum_{i=1}^{R}\sum_{j=1}^{C} E_{ij}(2\delta_{ij} + \delta_{ij}^2) = 2\sum_{i=1}^{R}\sum_{j=1}^{C}(O_{ij} - E_{ij}) + \sum_{i=1}^{R}\sum_{j=1}^{C}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} \\
&= \sum_{i=1}^{R}\sum_{j=1}^{C}\frac{(O_{ij}-E_{ij})^2}{E_{ij}},
\end{aligned}$$

where we use the facts that δ_{ij}^3 is negligible, the O_{ij}'s clearly sum to n, and:

$$\sum_{i=1}^{R}\sum_{j=1}^{C} E_{ij} = \frac{1}{n}\sum_{i=1}^{R}\sum_{j=1}^{C} Y_{i\bullet}Y_{\bullet j} = \frac{1}{n}\sum_{i=1}^{R} Y_{i\bullet}\sum_{j=1}^{C} Y_{\bullet j} = \frac{1}{n}(n \times n) = n,$$

which implies that $\sum_{i=1}^{R}\sum_{j=1}^{C}(O_{ij} - E_{ij}) = 0$.

The quantity,

$$X^2 = \sum_{i=1}^{R}\sum_{j=1}^{C} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

is called the *Pearson Chi-Squared Statistic*, and is the more common test statistic used to examine the *goodness-of-fit* of the independence model. In particular, we will reject the null hypothesis of independence of the two variables under study at significance level α if $X^2 \geq \chi^2_{(R-1)(C-1)}(1-\alpha)$.

In fact, the Pearson Chi-Squared Statistic can be used to test the *goodness-of-fit* of quite general hypothesised structures for the π_{ij}'s. All that is required is a determination of the appropriate E_{ij} values under the hypothesis of interest, and of course the appropriate degrees of freedom (derived from difference in the number of effective parameters associated with the hypothesis of interest and $RC - 1$, which is the effective number of degrees of freedom from the saturated model). Typically, it is the determination of the number of effective parameters of the hypothesis under study that is the most difficult task, but as a general rule, the effective number of parameters corresponds to the minimum number of π_{ij} values which, if known, determine the values of the rest of the π_{ij}'s. Once these quantities are calculated, the usual comparison of the Pearson chi-squared statistic to the critical values of a chi-squared distribution with the appropriate degrees of freedom can be made.

A note of caution is in order, since the distribution of the Pearson chi-squared statistic (as well as the distribution of the deviance statistic) is only approximately chi-squared, and this approximation is only truly reliable when a large majority (say, 80%) of the E_{ij} values are larger than 5. If there are a large number of E_{ij} values which are less than 5, one option is to combine categories within the variables to make the new E_{ij}'s larger. Alternatively, rows or columns with only a few observations (which is where the small E_{ij}'s will tend to lie) may be simply ignored in the analysis under the assumption that they are not a significant sector of the population. Of course, either of these approaches changes the direct interpretation of the π_{ij}'s and thus of the hypothesis test, so care must be taken not to introduce a bias into the analysis.

As further justification for the Pearson test statistic, note that if the data are assumed to come from a Poisson structure (recall that the multinomial distribution can be seen as arising from conditioning arguments based on the Poisson distribution), then the Pearson residuals would be:

$$r_{ij} = \frac{Y_{ij} - \tilde{\mu}_{ij}}{\sqrt{V(\tilde{\mu}_{ij})}} = \frac{O_{ij} - E_{ij}}{\sqrt{E_{ij}}},$$

since $V(\mu) = \mu$ for the Poisson distribution. Thus, the Pearson chi-squared statistic is just the sum of the squares of these Pearson residuals:

$$X^2 = \sum_{i=1}^{R}\sum_{j=1}^{C} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \sum_{i=1}^{R}\sum_{j=1}^{C} r_{ij}^2.$$

Example 1 (Continued): It is not difficult to construct the Pearson chi-squared statistic directly:

```
> views <- read.table("Sxviews.txt",header=T)
> views
```

```
                    East.Coast South Midwest West.Coast Outside.U.S.
      Very.Acceptable         82   201      30        169           13
  Somewhat.Acceptable         72   251      20         96            7
Somewhat.Unacceptable         49   122      26         36            4
    Very.Unacceptable         36   149      13         28            3
> rowtots <- apply(views,1,sum)
> rowtots
 Very.Acceptable Somewhat.Acceptable Somewhat.Unacceptable Very.Unacceptable
             495                  446                   237               229
> coltots <- apply(views,2,sum)
> coltots
 East.Coast South Midwest West.Coast Outside.U.S.
        239   723      89        329           27
> eij <- rowtots%*%t(coltots)/sum(rowtots)
> eij
     East.Coast    South  Midwest West.Coast Outside.U.S.
[1,]   84.08316 254.3603 31.31130  115.74627     9.498934
[2,]   75.75977 229.1812 28.21180  104.28856     8.558635
[3,]   40.25800 121.7846 14.99147   55.41791     4.547974
[4,]   38.89908 117.6738 14.48543   53.54726     4.394456
> oij <- as.matrix(views)
> prsnchi2 <- sum(((oij-eij)^2)/eij)
> c(prsnchi2,1-pchisq(prsnchi2,(4-1)*(5-1)))
[1] 8.088038e+01 2.802980e-12
```

reject H_0

So, clearly, independence is not a reasonable assumption for these data. Note that we can also "trick" the `glm()` function into giving us these values (actually, it is not a trick at all, but based again on the strong conditional relationship between the multinomial distribution and the Poisson distribution):

```
> yij <- as.vector(oij)
> rfact <- rep(1:4,5)
> cfact <- rep(1:5,rep(4,5))
> views.glm <- glm(yij ~ factor(rfact)+factor(cfact),family=poisson)
> sum(residuals(views.glm,"pearson")^2)
[1] 80.8802
> views.glm1 <- glm(yij ~ factor(rfact)*factor(cfact),family=poisson)
> anova(views.glm1)
Analysis of Deviance Table

Poisson model

Response:  yij

Terms added sequentially (first to last)

                           Df Deviance Resid. Df Resid. Dev
                         NULL                    19   1304.218
               factor(rfact)  3  166.243        16   1137.975
               factor(cfact)  4 1057.785        12     80.190
factor(rfact):factor(cfact) 12   80.190         0      0.000
```

Sum of squared Pearson residuals

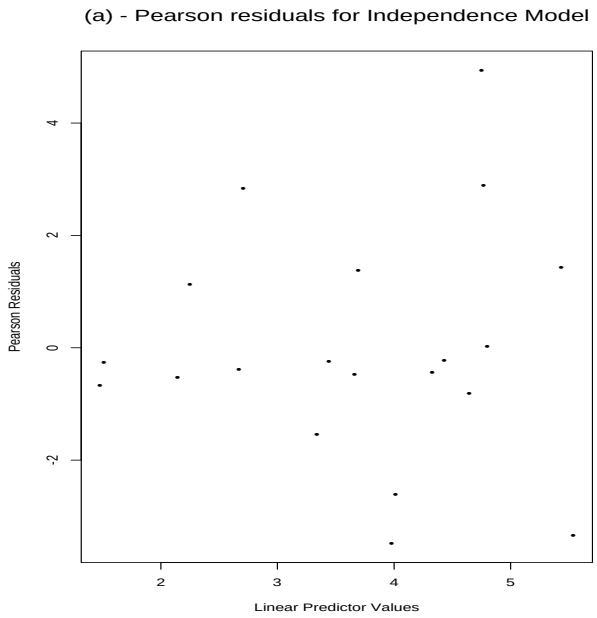
\$ Pearson chi-squared statistic

Note that the minor difference in the sum of the squared Pearson residuals and the Pearson chi-squared statistic is just due to rounding error as they are measuring the same thing. On the other hand the difference between the deviance statistic for the interaction terms and the Pearson chi-squared statistic is due to the fact that these are two different statistics and are only approximately equal to one another. Also, notice that the appropriate deviance statistic for testing independence is also the value of the residual deviance for the model fit assuming independence (i.e., with the "+" instead of the "*"), since the residual deviance for the saturated model is 0, as it is for any saturated model. So, the second model fit was not strictly necessary.

Now, once we have concluded that there is an association between the two variables under study, it remains to investigate what this relationship is. Previously, for GLMs, we started by examining the parameter estimates; however, in the present case that is not overly enlightening, partly due to the fact that some of the parameters may be just *constraint parameters*. It turns out that the most straightforward method of investigating the relationship between the variables is by investigating the Pearson residuals. Recall that the Pearson chi-squared statistic is just the sum of the squares of the Pearson residuals, and thus any cell with a large Pearson residual (i.e., large in absolute value) is contributing a large amount to the value of X^2 and it is here that we should investigate the relationship. Note that previously we simply called points with large residual values "outliers", but here it is exactly the outliers which tell us the most information (which again justifies our previous claims that outliers should be examined carefully rather than just being summarily discarded). For example, if a particular cell has a large positive residual, this means that $O_{ij} \gg E_{ij}$, so that the observed count is larger than would be expected if the variables were independent. Such cells indicate that individuals in the j^{th} column are more likely to be in the i^{th} row than individuals in the other columns. Similarly, large negative residuals indicate that $O_{ij} \ll E_{ij}$, implying that individuals in the j^{th} column are less likely to be in the i^{th} row than individuals in the other columns.

Example 1 (Continued): For this example, we can examine the Pearson residual values to determine how the two variables are related. To do so, we will use two different plots. The first is just a plot of the Pearson residuals versus the linear predictor values from the "auxiliary" Poisson GLM, which just gives us an idea of how large the residuals are. The second plot is more three dimensional in nature and shows which cells have large associated residual values, enabling us to graphically see any spacial relationships in where the large Pearson residuals lie.

```
> round((oij-eij)/sqrt(eij),3)
                     East.Coast  South Midwest West.Coast Outside.U.S.
      Very.Acceptable     -0.227 -3.346  -0.234      4.950        1.136
  Somewhat.Acceptable     -0.432  1.441  -1.546     -0.812       -0.533
Somewhat.Unacceptable      1.378  0.020   2.843     -2.608       -0.257
    Very.Unacceptable     -0.465  2.888  -0.390     -3.491       -0.665
> plot(views.glm$linear.predictor,residuals(views.glm,"pearson"),
+   xlab="Linear Predictor Values",ylab="Pearson Residuals",
+   main="(a) - Pearson residuals for Independence Model")
> tmp <- list(grand=0,row=1:nrow(views),col=(1:ncol(views))*1.4,
+   resid=residuals(views.glm,"pearson"))
> plotfit(tmp,rowlab=dimnames(views)[[1]],collab=dimnames(views)[[2]],c=3,
+   main="(b) - Pearson residuals by Cell",yaxt="n")
```



Note that the argument `c=` tells the graphical function `plotfit()` to only draw vertical lines for residuals with absolute values larger than the specified amount. From these plots, we can see that the lack of independence arises primarily from the fact that students raised in the Southern U.S. (a highly conservative region) were much less likely to view premarital sex as "very acceptable", while students raised on the U.S. West Coast (a highly liberal region) were much more likely to view premarital sex as "very acceptable" and much less likely to view it as "very unacceptable". Equivalently, the results may be interpreted as demonstrating that students who believe that premarital sex is "very acceptable" are much more likely to have been raised on the West Coast and much less likely to have been raised in the South, while students who feel that premarital sex is "very unacceptable" are much less likely to have been raised on the West Coast.

We now shift our attention to the case in which the data was sampled with fixed column totals. The log-likelihood for this case is derived from the product multinomial probability function:

$$\begin{aligned} l(\mu|Y_{\bullet 1},\ldots,Y_{\bullet C}) &= \sum_{j=1}^{C}\left\{\log Y_{\bullet j}! + \sum_{i=1}^{R}\left(Y_{ij}\log\pi_{ij} - \log Y_{ij}!\right)\right\} \\ &= \sum_{j=1}^{C}\left[\log Y_{\bullet j}! + \sum_{i=1}^{R}\left\{Y_{ij}\log\left(\frac{\mu_{ij}}{Y_{\bullet j}}\right) - \log Y_{ij}!\right\}\right] \\ &= \sum_{i=1}^{R}\sum_{j=1}^{C} Y_{ij}\log\mu_{ij} + \sum_{j=1}^{C}\left\{\log Y_{\bullet j}! - \sum_{i=1}^{R}\left(Y_{ij}\log Y_{\bullet j} + \log Y_{ij}!\right)\right\} \\ &= \sum_{i=1}^{R}\sum_{j=1}^{C} Y_{ij}\log\mu_{ij} + d(Y), \end{aligned}$$

which is nearly identical to the log-likelihood for the case of fixed total sampling; the only difference being that the $d(Y)$ function has a slightly different form, but this is unimportant as we saw. Moreover, the fitted values for the saturated model are clearly still the same, so that $\hat{\mu}_{ij} = Y_{ij} = O_{ij}$. Also, under the model of homogeneous column distributions we have $Y_{\bullet j}\pi_{ij} = Y_{\bullet j}\pi_{i\bullet}$, so that the fitted values under this model can be readily seen to be:

$$Y_{\bullet j}\hat{\pi}_{i\bullet} = \frac{Y_{i\bullet}Y_{\bullet j}}{n} = \tilde{\mu}_{ij} = E_{ij}.$$

which are identical to the fitted values in the case of the independence model for sampling with a fixed total only (which is why we maintain the same notation).

Once we see these similarities, we can immediately infer that the appropriate test of homogeneity is exactly the same as the one for independence. In other words, procedurally the two hypothesis tests are identical and we will reject the null hypothesis of homogeneity at significance level α if

$$X^2 \geq \chi^2_{(R-1)(C-1)}(1-\alpha).$$

Similarly, the examination of the Pearson residuals to investigate any inhomogeneity in the column distributions is carried out exactly as it was for the case of independence testing. So, while the philosophy of the two approaches is somewhat different, the analysis techniques involved are identical. In some sense, the only real distinction between the two approaches is that the column proportions, $\hat{\pi}_{\bullet j}$, are meaningful estimates of the proportions of individuals in the population under study which fall into each column category when the sampling scheme has only a fixed total, while they are not meaningful estimates in the case of fixed column totals, precisely because the sampling design has fixed these values rather than letting the population itself dictate the number of individuals who will fall into each column category. Indeed, as the next example shows, it is often the case that the column proportions are not of any interest at all, but are merely structural components of the data.

Example 2 - Cheese Tasting: Four different cheese additives (designated as A, B, C and D) were examined for flavor by 52 different tasters. The flavor was judged on the so-called hedonic scale which ranges from I ('strong dislike') to IX ('excellent taste'), and each judge rated every additive. The data is tabulated in a contingency table indicating the number of judges who rated each cheese at each possible scale value. Clearly this is a case where the column totals are fixed by the sampling scheme, since we have ensured that each judge rates each cheese additive, so the column totals must be simply the total number of tasters.

```
> chs <- read.table("Cheese.txt",header=T)
> attach(chs)
> chs
      A  B  C  D
   I  0  6  1  0
  II  0  9  1  0
 III  1 12  6  0
  IV  7 11  8  1
   V  8  7 23  3
  VI  8  6  7  7
 VII 19  1  5 14
VIII  8  0  1 16
  IX  1  0  0 11
> rtot <- apply(chs,1,sum)
> rtot
 I II III IV  V VI VII VIII IX
 7 10  19 27 41 28  39   25 12
> ctot <- apply(chs,2,sum)
> ctot
  A  B  C  D
 52 52 52 52
> eij <- rtot%*%t(ctot)/sum(rtot)
> eij
```

```
          A     B     C     D
[1,]  1.75  1.75  1.75  1.75
[2,]  2.50  2.50  2.50  2.50
[3,]  4.75  4.75  4.75  4.75
[4,]  6.75  6.75  6.75  6.75
[5,] 10.25 10.25 10.25 10.25
[6,]  7.00  7.00  7.00  7.00
[7,]  9.75  9.75  9.75  9.75
[8,]  6.25  6.25  6.25  6.25
[9,]  3.00  3.00  3.00  3.00
```

Expected # of counts for each rating (row)

Now, from the table of expected values, we see that the first three and the last rows have E_{ij} values less than five (which is 16/36=44.4%). To remedy this, we will combine the first three and the last two rows together (of course, this changes the structure of the data, but it should not do so too dramatically in this case, since so few counts are in these rows to begin with, and thus we can still meaningfully test which additive is the best liked).

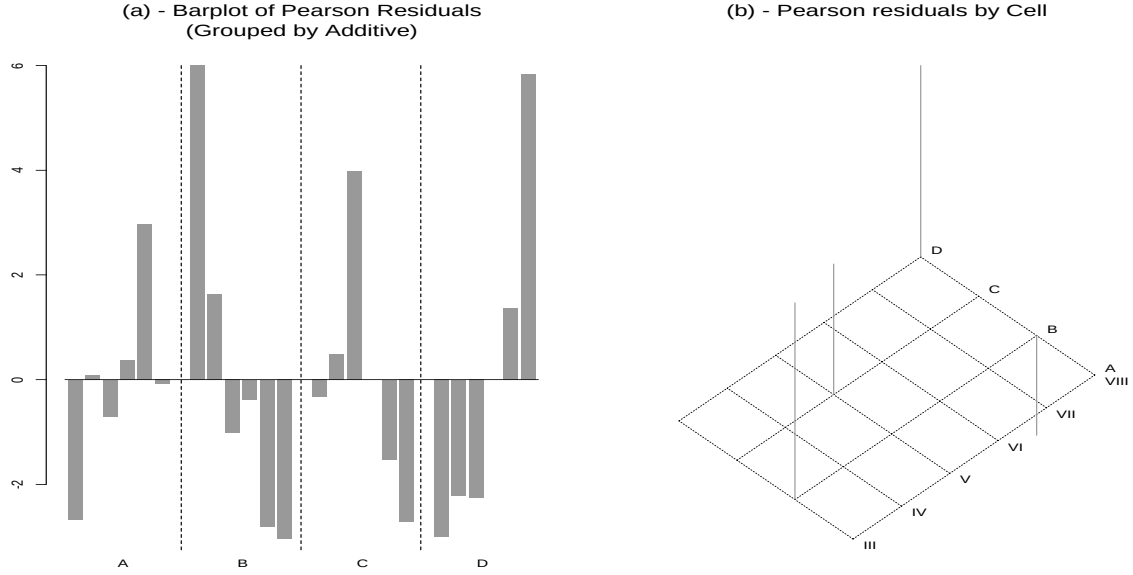
```
> chs1 <- chs[3:8,]
> chs1[1,] <- chs1[1,]+chs[1,]+chs[2,]
> chs1[6,] <- chs1[6,]+chs[9,]
> rtot1 <- apply(chs1,1,sum)
> eij1 <- rtot1%*%t(ctot)/sum(rtot1)
> eij1
          A     B     C     D
[1,]  9.00  9.00  9.00  9.00
[2,]  6.75  6.75  6.75  6.75
[3,] 10.25 10.25 10.25 10.25
[4,]  7.00  7.00  7.00  7.00
[5,]  9.75  9.75  9.75  9.75
[6,]  9.25  9.25  9.25  9.25
> oij1 <- as.matrix(chs1)
> rij1 <- (oij1-eij1)/sqrt(eij1)
> c(sum(rij1^2),1-pchisq(sum(rij1^2),(6-1)*(4-1)))
[1] 154.3055 > 0.0000
> rij1
                A          B          C         D
 III -2.66666667  6.0000000 -0.3333333 -3.000000
  IV  0.09622504  1.6358258  0.4811252 -2.213176
   V -0.70278193 -1.0151295  3.9824309 -2.264520
  VI  0.37796447 -0.3779645  0.0000000  0.000000
 VII  2.96237085 -2.8022427 -1.5212175  1.361089
VIII -0.08219949 -3.0413813 -2.7125833  5.836164
```

reject H_0

Examining the Pearson residuals using the above table and the plots below, we clearly see that the rejection of the null hypothesis of homogeneity is due to the fact that additive D is the most preferred, while additive B is the least preferred. Also, it appears that additive A is preferred over additive C, so the ordering of the additives in terms of taste is D, A, C, B, from best to worst.

```
> tmp <- c(rij1[,1],0,rij1[,2],0,rij1[,3],0,rij1[,4])
> barplot(tmp,main="(a) - Barplot of Pearson Residuals
```

```
Continue string:  (Grouped by Additive)")
> lines(c(8,8),c(6,-3.3),lty=3)
> lines(c(16.25,16.25),c(6,-3.3),lty=3)
> lines(c(24.5,24.5),c(6,-3.3),lty=3)
> text(4,-3.5,"A")
> text(12.25,-3.5,"B")
> text(20.5,-3.5,"C")
> text(28.75,-3.5,"D")
> tmp <- list(grand=0,row=1:nrow(chs1),col=(1:ncol(chs1))*1.2,
+   resid=as.vector(rij1))
> plotfit(tmp,rowlab=dimnames(chs1)[[1]],collab=dimnames(chs1)[[2]],c=3,
+   main="(b) - Pearson residuals by Cell",yaxt="n")
```



For completeness, we now examine the connection between contingency table analysis and logistic regression in the case where $R = 2$. We shall work under the scheme where the column totals are assumed to be fixed. It this case, the appropriate product multinomial distribution reduces to:

$$\begin{aligned}
f(\mu|Y_{\bullet 1}, \ldots, Y_{\bullet C}) &= \prod_{j=1}^{C} Y_{\bullet j}! \prod_{i=1}^{2} \frac{\pi_{ij}^{Y_{ij}}}{Y_{ij}!} \\
&= \prod_{j=1}^{C} Y_{\bullet j}! \frac{\pi_{1j}^{Y_{1j}}}{Y_{1j}!} \frac{\pi_{2j}^{Y_{2j}}}{Y_{2j}!} \\
&= \prod_{j=1}^{C} \frac{Y_{\bullet j}!}{Y_{1j}! Y_{2j}!} \pi_{1j}^{Y_{1j}} \pi_{2j}^{Y_{2j}} \\
&= \prod_{j=1}^{C} \frac{Y_{\bullet j}!}{Y_{1j}!(Y_{\bullet j} - Y_{1j})!} \pi_{1j}^{Y_{1j}} (1 - \pi_{1j})^{Y_{\bullet j} - Y_{1j}} \\
&= \prod_{j=1}^{C} Bin(Y_{\bullet j}, \pi_{1j}),
\end{aligned}$$

where we have used the fact $R = 2$ implies that $Y_{2j} = Y_{\bullet j} - Y_{1j}$ and, along with the constraints on the π_{ij}'s for this model (i.e., $\pi_{\bullet j} = 1$ for $1 \leq j \leq C$), also implies that $\pi_{2j} = 1 - \pi_{1j}$. Also, we

have employed the notation $Bin(Y_{\bullet j}, \pi_{1j})$ to designate the probability mass function for a binomial distribution with $Y_{\bullet j}$ trials and success probability π_{1j}. Thus, we can see that in this situation, the distribution being employed is exactly the same as if we were conducting a logistic regression with response variable $Y_{1j}/Y_{\bullet j}$ and a categorical predictor variable having C levels.

Now, the log-linear model approach to this problem starts from the link function:

$$\log \mu_{ij} = \log n + \tau_i + \alpha_j + \gamma_{ij} \Longrightarrow \log \pi_{ij} = \log\left(\frac{\mu_{ij}}{Y_{\bullet j}}\right) = \log \mu_{ij} - \log Y_{\bullet j} = \tau_i + \alpha_j + \gamma_{ij} + \log\left(\frac{n}{Y_{\bullet j}}\right).$$

Thus, for the two rows separately, we have

$$\log \pi_{1j} = \tau_1 + \alpha_j + \gamma_{1j} + \log\left(\frac{n}{Y_{\bullet j}}\right) \qquad \text{and} \qquad \log \pi_{2j} = \tau_2 + \alpha_j + \gamma_{2j} + \log\left(\frac{n}{Y_{\bullet j}}\right).$$

Subtracting these two equations shows that the log-linear model is equivalent to the model

$$\log\left(\frac{\pi_{1j}}{\pi_{2j}}\right) = (\tau_1 - \tau_2) + (\gamma_{1j} - \gamma_{2j}).$$

Note that the only parameters which are missing from this model are the $\log\left(\frac{n}{Y_{\bullet j}}\right)$ and α_j's which are precisely the constraint parameters and thus were only structural in nature, so that there removal from the model does not change its interpretation at all. Moreover, since $\pi_{2j} = 1 - \pi_{1j}$ we see that the above model is actually logistic regression:

$$\log\left(\frac{\pi_{1j}}{1 - \pi_{1j}}\right) = (\tau_1 - \tau_2) + (\gamma_{1j} - \gamma_{2j}) = \beta_0 + \beta_j,$$

where $\beta_0 = (\tau_1 - \tau_2)$ and $\beta_j = (\gamma_{1j} - \gamma_{2j})$. So, we see that testing whether the interaction term (i.e., the γ's) are unimportant in the model (which is the test for homogeneity) is equivalent to testing for the significance of a categorical predictor in a logistic regression.

Of course, the logistic regression approach has the advantage that it can handle ordinal or continuous covariates and the contingency table analysis we have developed so far cannot (except by simply ignoring any ordering information in the predictor values). So, it is generally best to use the logistic regression approach whenever possible, which means whenever there are only two rows in our contingency table. However, this leads us nicely into our next topic, which is how to handle ordinal variables in contingency table analyses.

V. Two-way Contingency Table Analysis with Ordinal Variables

In this section, we will give a brief discussion of how one might take account of the ordering information in ordinal variables. This is a subject in which much research is still being done, and definitive answers are not yet possible on what the optimal methods for analysing such data are. However, we can give some basic results which work well in limited circumstances. For this discussion, we will assume that the data have been gathered under a sampling scheme which has fixed column totals, and we will further assume that the ordinal variable is the row variable (often thought of as the response variable in this situation), while the column variable is nominal. This is the most frequent situation in practice, though ordinal variables do occur in other settings and methods for analysis are available (they are just too complicated for inclusion in this course).

The most straightforward way to incorporate the ordering of the ordinal variable is to model the *cumulative probabilities*:

$$\theta_{ij} = \sum_{k=1}^{i} \pi_{kj}.$$

Note that $\theta_{1j} = \pi_{1j}$ and $\theta_{Rj} = \pi_{\bullet j} = 1$, under the fixed column totals sampling scheme. Also, notice that this approach only makes sense for ordinal variables, since it is precisely the ordering of the categories which determines the appropriate sequence in which the π_{ij}'s are to be added. In general then, we might wish to model the θ_{ij}'s as:

$$g(\theta_{ij}) = \mu + \tau_i + \alpha_j,$$

for some "link" function g, such as the logistic function or the complementary log-log function, or just the logarithmic function itself [TECHNICAL NOTE: As usual, we need a constraint to make the above model "identifiable", since in its current form we can simply add a constant c to all the τ's and subtract the same constant from all the α's and arrive at the same values for $g(\theta_{ij})$ for two different sets of τ and α values.] Unfortunately, the fact that we call g a link function does not mean that the model will necessarily be a GLM, and indeed in this case it cannot be, since θ_{ij} is not the expectation of any individual observation. Nonetheless, we can write the product multinomial log-likelihood function in terms of the θ_{ij}'s as:

$$\begin{aligned} l(\theta|Y_{\bullet 1}, \ldots, Y_{\bullet C}) &= \sum_{i=1}^{R}\sum_{j=1}^{C} Y_{ij} \log \mu_{ij} + d(Y) \\ &= \sum_{i=1}^{R}\sum_{j=1}^{C} Y_{ij} \log(Y_{\bullet j}\pi_{ij}) + d(Y) \\ &= \sum_{i=1}^{R}\sum_{j=1}^{C} Y_{ij} (\log \pi_{ij} + \log Y_{\bullet j}) + d(Y) \\ &= \sum_{j=1}^{C} Y_{1j} \log \theta_{1j} + \sum_{i=2}^{R}\sum_{j=1}^{C} Y_{ij} \log(\theta_{ij} - \theta_{i-1,j}) + d_1(Y), \end{aligned}$$

where d_1 is just another unimportant function appropriately defined. Now, the fact that the model we wish to examine is not a GLM simply means that it is more difficult to find the MLEs (i.e., we cannot use *S-Plus* directly), but the principle of maximum likelihood estimation remains the same. In other words, if we substitute $g^{-1}(\tau_i + \alpha_j)$ for θ_{ij} in the log-likelihood, we can then find the MLEs $\hat{\tau}_i$ and $\hat{\alpha}_j$ by differentiating and solving the *score equations* (i.e., the derivatives set equal to zero).

Suppose we believe that the distribution of the row variable is homogeneous within each of the levels of the column variable. Then, we know that $\pi_{ij} = \pi_{i\bullet}$, and this implies that under the assumption of homogeneity:

$$\theta_{ij} = \sum_{k=1}^{i} \pi_{kj} = \sum_{k=1}^{i} \pi_{k\bullet},$$

which is the same for all values of j. In other words, the assumption of homogeneity can be seen in terms of the model structure as being the assumption that the α_j's are all equal to zero. Under this scheme, it can be shown (with a bit of painful algebra) that the MLEs for the θ_{ij}'s are given by:

$$\tilde{\theta}_{ij} = \sum_{k=1}^{i} \frac{Y_{k\bullet}}{n}.$$

So, it only remains to find the MLEs under the larger model, $\hat{\theta}_{ij}$, and then we can appeal to the drop in deviance statistic $D^\star(\tilde{\theta}_{ij}, \hat{\theta}_{ij})$, to test our homogeneity model.

Unfortunately, in general finding the $\hat{\theta}_{ij}$'s is a very difficult proposition and requires iterative schemes similar to the $IRLS$ algorithm which are beyond the scope of this course. There is one

situation, however, in which we can find the MLEs using the `glm()` function of *S-Plus* in an indirect way. This is the case where g is the complementary log-log function, and in this case our model is generally referred to as the *proportional hazards model*, which has many applications in biostatistics and survival analysis (and indeed, it is from these fields that the model gets its name, as it is generally used to model how the chance of some detrimental outcome is affected by exposure to various levels of some physical or biological hazard). It turns out (though the algebra required to show this fact is prodigious and thus omitted) that we can find the MLEs of the α_j's (and indeed perform the entire analysis) for the proportional hazards model by fitting a binomial GLM to a transformation of the observed counts. Specifically, we define

$$\begin{aligned} Z_{1j} &= \frac{Y_{1j}}{Y_{\bullet j}} \\ Z_{2j} &= \frac{Y_{2j}}{Y_{\bullet j} - Y_{1j}} \\ &\vdots \\ Z_{ij} &= \frac{Y_{ij}}{Y_{\bullet j} - \sum_{k=1}^{i-1} Y_{kj}} \\ &\vdots \\ Z_{R-1,j} &= \frac{Y_{R-1,j}}{Y_{\bullet j} - \sum_{k=1}^{R-2} Y_{kj}}, \end{aligned}$$

and we fit a complementary log-log binomial GLM using the Z_{ij}'s (note that there are now $(R-1)C$ of these values, since clearly the logical extension of the above definitions show that $Z_{R,j} = 1$ and are thus unimportant information) as the response values, appropriate row and column indicators z_1 up to z_{R-2} and w_1 up to w_{C-1} (i.e., z_i is an indicator of whether or not the corresponding response is in the $(i+1)^{\text{st}}$ row and w_i is an indicator of whether the corresponding response is in the $(i+1)^{\text{st}}$ column, implying that we have used a "baseline" constraint system here, so that by assumption $\alpha_1 = 0$) as the predictor variables and additional weights equal to the denominators in the definition of the Z_{ij}'s (NOTE: If we imagine counting individuals in the j^{th} column sequentially, first determining how many are in the first category of the row variable and then counting how many are in the second category and so on, then the Z_{ij} values are just the proportion of all individuals in the j^{th} column who are yet to be counted which happen fall into the i^{th} level of the row variable. In other words, the Z_{ij}'s are just the proportion of individuals not yet counted who *will* be counted at the next tabulation.) The test for homogeneity can then be performed using the analysis of deviance table for this GLM to assess whether the column indicators are significant in the model. Furthermore, it can be shown that the $\hat{\theta}_{ij}$'s can be calculated from the fitted values of this GLM as:

$$\begin{aligned} \hat{\theta}_{1j} &= \hat{Z}_{1j} \\ \hat{\theta}_{ij} &= \hat{Z}_{ij} + \hat{\theta}_{i-1,j}(1 - \hat{Z}_{ij}) \qquad 1 < i < R \end{aligned}$$

where the $\hat{Z}_{ij}$'s are the fitted values from the complementary log-log binomial GLM.

Example 2 (Continued): In the cheese tasting example, the row categories are levels on the hedonic rating scale and thus clearly give rise to an ordinal variable. So, we can test for homogeneity using the proportional hazards model as:

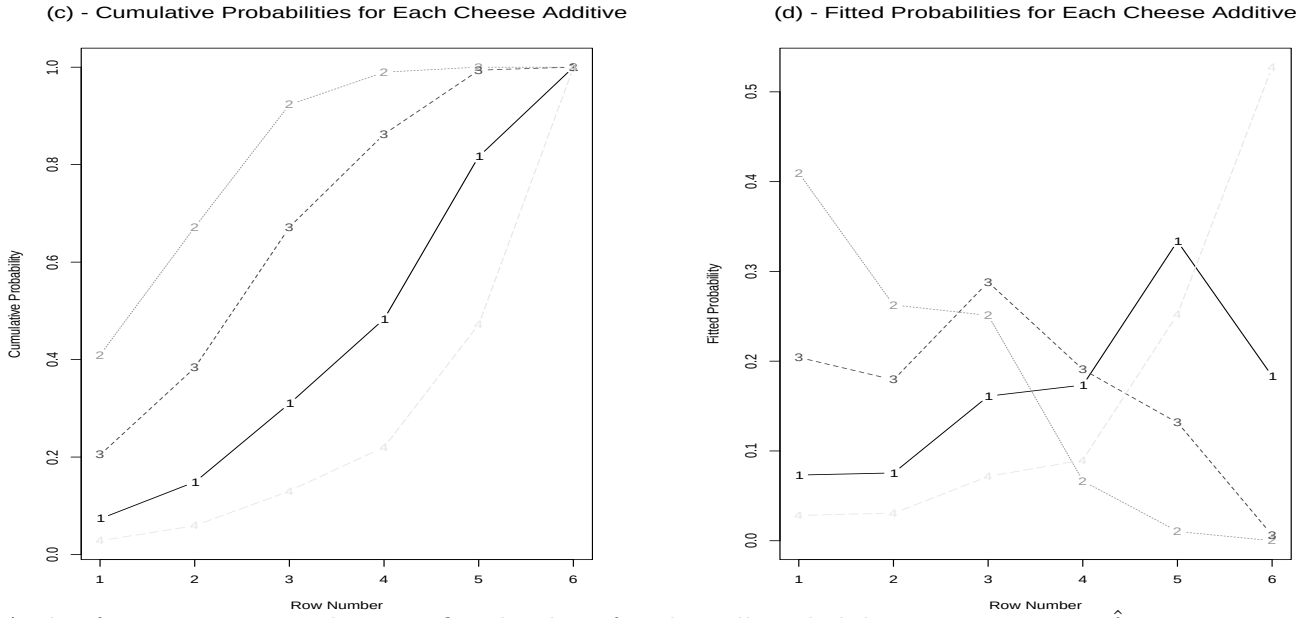
```
> nij <- rbind(ctot,ctot,ctot,ctot,ctot)
> nij[-1,] <- nij[-1,] - apply(chs1,2,cumsum)[1:4,]
> zij <- chs1[1:5,]/nij
```

```
> rfct <- cbind(1:5,1:5,1:5,1:5)
> cfct <- rbind(1:4,1:4,1:4,1:4,1:4)
> prp <- as.vector(zij)
> wgt <- as.vector(nij)
> rfct <- as.vector(rfct)
> cfct <- as.vector(cfct)
> rind1 <- ifelse(rfct==2,1,0)
> rind2 <- ifelse(rfct==3,1,0)
> rind3 <- ifelse(rfct==4,1,0)
> rind4 <- ifelse(rfct==5,1,0)
> rinds <- cbind(rind1,rind2,rind3,rind4)
> cind1 <- ifelse(cfct==2,1,0)
> cind2 <- ifelse(cfct==3,1,0)
> cind3 <- ifelse(cfct==4,1,0)
> cinds <- cbind(cind1,cind2,cind3)
> chs1.glm <- glm(prp ~ rinds+cinds,family=binomial(link=cloglog),weights=wgt)
> anova(chs1.glm)
Analysis of Deviance Table

Binomial model

Response:  prp

Terms added sequentially (first to last)
      Df Deviance Resid. Df Resid. Dev
 NULL                    19   201.9414
rinds  4  40.8020        15   161.1393
cinds  3 137.2932        12    23.8461
> 1-pchisq(137.2932,3)
[1] 0
```

So, clearly the column effect is significant, meaning that the columns are not homogeneous. It remains then to interpret the coefficient estimates $\hat{\alpha}_j$ for the larger model. To do this, we note that for a fixed value of τ_i, increasing α_j implies that $Z_{ij} = 1 - \exp(-\exp(\tau_i + \alpha_j))$ increases as well. So, the larger the value of $\hat{\alpha}_j$, the more quickly the θ_{ij}'s increase towards unity (as the Z_{ij}'s are just the proportion of individuals not yet counted in the j^{th} column who will be counted in the i^{th} row.). This means that columns with a large value of $\hat{\alpha}_j$ have a tendency to have most of their counts in the *initial* row categories. For our cheese example, this means that the larger the value of $\hat{\alpha}_j$, the *less* preferred the additive.

```
> summary(chs1.glm)$coef[length(coef(chs1.glm))+(-2:0),]
                Value Std. Error    t value
cindscind1  1.9357607  0.2473345  7.826488
cindscind2  1.1007412  0.2309313  4.766530
cindscind3 -0.9755664  0.2545834 -3.832010
> hatzij <- matrix(fitted(chs1.glm),ncol=4)
> cumprbs <- matrix(0,5,4)
> cumprbs[1,] <- hatzij[1,]
> for(i in 2:5) {
+  cumprbs[i,] <- hatzij[i,]+(cumprbs[i-1,]*(1-hatzij[i,]))
+ }
```

```
> cumprbs
           [,1]      [,2]      [,3]       [,4]
[1,] 0.07320397 0.4094959 0.2043146 0.02825183
[2,] 0.14864282 0.6721141 0.3835632 0.05886130
[3,] 0.30959857 0.9232488 0.6716959 0.13034951
[4,] 0.48299500 0.9896554 0.8623892 0.22018139
[5,] 0.81662557 0.9999921 0.9939003 0.47241268
> matplot(1:6,rbind(cumprbs,1),type="b",ylab="Cumulative Probability",xlab=
+  "Row Number",main="(c) - Cumulative Probabilities for Each Cheese Additive")
```



And, of course, we can then get fitted values for the cell probabilities, since $\hat{\pi}_{1j} = \hat{\theta}_{1j}$, $\hat{\pi}_{Rj} = 1 - \hat{\theta}_{R-1,j}$ and $\hat{\pi}_{ij} = \hat{\theta}_{ij} - \hat{\theta}_{i-1,j}$ for $1 < i < R$:

```
> fitprbs <- rbind(cumprbs,1)-rbind(0,cumprbs)
> fitprbs
           [,1]         [,2]        [,3]       [,4]
[1,] 0.07320397 4.094959e-01 0.204314568 0.02825183
[2,] 0.07543885 2.626182e-01 0.179248601 0.03060947
[3,] 0.16095575 2.511347e-01 0.288132704 0.07148821
[4,] 0.17339643 6.640655e-02 0.190693350 0.08983188
[5,] 0.33363057 1.033678e-02 0.131511105 0.25223130
[6,] 0.18337443 7.860378e-06 0.006099672 0.52758732
> matplot(1:6,fitprbs,type="b",ylab="Fitted Probability",xlab="Row Number",
+   main="(d) - Fitted Probabilities for Each Cheese Additive")
```

So, as we saw from our previous analysis, the ordering of the additives from most to least preferred is still D, A, C, B. However, the current analysis shows us this a bit more quantitatively, and is more appropriate to the data. Furthermore, the interpretation of our analysis is generally clearer when we look at the plot of the cumulative probabilities than if we look at the fitted cell probabilities themselves (though, in this case the results are reasonably clear from both plots due to the strong preferences displayed by the tasters).

As a final note, we point out that we can assess whether the proportional hazards model itself seems reasonable for these data by examining its residual deviance, and comparing it to an appropriate chi-squared distribution. This is simply a check on the assumption that $\phi = 1$ for a

binomial GLM, since the residual deviance divided by its degrees of freedom is just an estimate, $\hat{\phi}$, of the dispersion. In this case, the residual deviance is 23.8461 with 12 degrees of freedom and:

```
> 1-pchisq(23.8461,12)
[1] 0.02134372
```

So, perhaps there is an overdispersion or model misspecification problem indicated here. However, the methods to correct this problem are outside the scope of this course. Recall however, that we chose the proportional hazards model here purely for mathematical and computational convenience (so that we could use the `glm()` function in *S-Plus*), and perhaps a different model structure would be more appropriate (though of course fitting such models can be quite complicated, and is again outside the scope of this course). Also, as a technical note, we may need to test whether the dispersion estimate is too small compared to one. This rarely ever happens in practice, and when it does is generally attributable to an overfit model to a small contingency table (i.e., few rows and columns). However, there is a quantitative measure we can use to formally test whether our residual deviance value is too small. In fact, it amounts to exactly the same procedure as we used above in testing whether the value was too large, except that we examine the opposite "tail" of the chi-squared distribution. Thus, we will accept the proportional hazards model as adequate as long the observed residual deviance for the model is between the 5% and 95% quantiles of the appropriate chi-squared distribution. (NOTE: Technically, this means that the observed p-value for the residual deviance value should lie between 0.05 and 0.95 for the model to be deemed adequate.)

VI. Multi-way Contingency Table Analysis

In this section, we extend our analysis to count data in a multi-way contingency table. In particular, we will examine three-way contingency tables and the logical extension to multi-way tables is then not overly difficult. So, suppose that we have count data denoted as Y_{ijk} (or equivalently as O_{ijk}), where the subscript notation indicates to which levels of each of the three categorical variables the count corresponds. Also, we will work under the assumptions of a sampling scheme where only the total sample size, n, is fixed and the categorical variables under study are all nominal. Then, the appropriate multinomial distribution function for $Y = (Y_{111}, \ldots, Y_{IJK})$ is:

$$f(Y|n) = n! \prod_{i=1}^{I} \prod_{j=1}^{J} \prod_{k=1}^{K} \frac{\pi_{ijk}^{Y_{ijk}}}{Y_{ijk}!},$$

where π_{ijk} is just the probability that a sampled individual contributes to the $(i,j,k)^{\text{th}}$ cell of the three-way contingency table, and we now denote the total number of levels for each of the three categorical variables as I, J and K, respectively. This leads to the log-likelihood:

$$\begin{aligned}
l(\mu|n) &= \log n! + \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\{Y_{ijk}\log(\pi_{ijk}) - \log Y_{ijk}!\} \\
&= \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} Y_{ijk}\log\left(\frac{\mu_{ijk}}{n}\right) + \log n! - \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\log Y_{ijk}! \\
&= \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} Y_{ijk}\log\mu_{ijk} + \log n! - \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\{\log Y_{ijk}! + Y_{ijk}\log n\} \\
&= \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} Y_{ijk}\log\mu_{ijk} + d(Y),
\end{aligned}$$

where we have taken advantage of the (intuitively clear) fact that

$$E(Y_{ijk}) = \mu_{ijk} = n\pi_{ijk}.$$

In direct analogy to the analysis of the two-way contingency table, we see that the MLEs under the saturated model are just:

$$\hat{\pi}_{ijk} = \frac{Y_{ijk}}{n} \qquad \Longrightarrow \qquad \hat{\mu}_{ijk} = Y_{ijk} = O_{ijk}.$$

Furthermore, under the assumption of independence of the three variables under study, we have $\pi_{ijk} = \pi_{i\bullet\bullet}\pi_{\bullet j\bullet}\pi_{\bullet\bullet k}$, where we have made the obvious notational definitions:

$$\pi_{i\bullet\bullet} = \sum_{j=1}^{J}\sum_{k=1}^{K}\pi_{ijk}; \qquad \pi_{\bullet j\bullet} = \sum_{i=1}^{I}\sum_{k=1}^{K}\pi_{ijk}; \qquad \pi_{\bullet\bullet k} = \sum_{i=1}^{I}\sum_{j=1}^{J}\pi_{ijk}.$$

So, the fitted values under the assumption of independence are just:

$$\tilde{\mu}_{ijk} = n\hat{\pi}_{i\bullet\bullet}\hat{\pi}_{\bullet j\bullet}\hat{\pi}_{\bullet\bullet k} = n\left(\frac{Y_{i\bullet\bullet}}{n}\right)\left(\frac{Y_{\bullet j\bullet}}{n}\right)\left(\frac{Y_{\bullet\bullet k}}{n}\right) = \frac{Y_{i\bullet\bullet}Y_{\bullet j\bullet}Y_{\bullet\bullet k}}{n^2} = E_{ijk},$$

where the "$\bullet$" subscript on the Y's indicates summation over the appropriate index value, so that the values in the above formula are the appropriate marginal totals of the three-way contingency table.

To test the hypothesis of independence, we can use the drop in deviance statistic which is:

$$D^{\star}(\tilde{\mu}, \hat{\mu}) = 2\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} O_{ijk}\log\left(\frac{O_{ijk}}{E_{ijk}}\right).$$

However, it is generally more common practice to use the Pearson chi-squared statistic which is:

$$X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\frac{(O_{ijk} - E_{ijk})^2}{E_{ijk}}.$$

Both of these statistics have approximate chi-squared distributions with $IJK - \{(I-1) + (J-1) + (K-1) + 1\} = IJK - I - J - K + 2$ degrees of freedom, so in general, we will reject the null hypothesis of independence at significance level α if

$$X^2 \geq \chi^2_{(IJK-I-J-K+2)}(1-\alpha).$$

df becomes

Example 3 - Insulin Dependence and Diabetes: A sample of 123 diabetic patients was cross-classified by family history of diabetes, their age at onset of the disease and whether or not they required daily insulin injections:

		Insulin Dependent	
Family History	Onset Age	Yes	No
Yes	< 45	6	1
	≥ 45	6	36
No	< 45	16	2
	≥ 45	8	48

To test whether these three variables are independent:

```
> yijk <- array(0,c(2,2,2),dimnames=list(c("<45yo",">45yo"),
+   c("ins.yes","ins.no"),c("hist.yes","hist.no")))
> yijk[,,1] <- matrix(c(6,6,1,36),ncol=2)
> yijk[,,2] <- matrix(c(16,8,2,48),ncol=2)
> yijk
, , hist.yes
      ins.yes ins.no
<45yo       6      1
>45yo       6     36
, , hist.no
      ins.yes ins.no
<45yo      16      2
>45yo       8     48
> agetot <- apply(yijk,1,sum)
> instot <- apply(yijk,2,sum)
> histtot <- apply(yijk,3,sum)
> tot <- sum(agetot)
> eijk <- array(0,c(2,2,2),dimnames=list(c("<45yo",">45yo"),
+   c("ins.yes","ins.no"),c("hist.yes","hist.no")))
> eijk[,,1] <- histtot[1]*agetot%*%t(instot)/(tot^2)
> eijk[,,2] <- histtot[2]*agetot%*%t(instot)/(tot^2)
> eijk
, , hist.yes
        ins.yes    ins.no
<45yo  2.914932  7.044418
>45yo 11.426532 27.614119
, , hist.no x
        ins.yes   ins.no
<45yo  4.402142 10.63851
>45yo 17.256395 41.70295
> prsd <- (yijk-eijk)/sqrt(eijk)
> c(sum(prsd^2),1-pchisq(sum(prsd^2),8-(2-1)-(2-1)-(2-1)-1))
[1] 5.706140e+01 1.201017e-11
```

So, clearly these three variables are not independent. Now, at this stage we must point out that unlike the case of two-way tables, there are several more possible hypotheses of interest in the multi-way contingency table case. For example, just because all three variables are not independent does not mean that some of the variables are not independent of others. We can test whether one of the variables is independent of the other two using a very similar chi-squared procedure to that given above. Suppose we wanted to test whether the first variable was independent of the other two. Then, under this assumption, the probabilities π_{ijk} can be written as:

$$\pi_{ijk} = \pi_{i\bullet\bullet}\pi_{\bullet jk},$$

and it is then intuitively clear that the appropriate fitted values are given by:

$$E_{ijk} = n\hat{\pi}_{i\bullet\bullet}\hat{\pi}_{\bullet jk} = n\left(\frac{Y_{i\bullet\bullet}}{n}\right)\left(\frac{Y_{\bullet jk}}{n}\right) = \frac{Y_{i\bullet\bullet}Y_{\bullet jk}}{n}.$$

Using these expected values in the Pearson chi-squared statistic, we can then test the hypothesis of independence of the first variable from the other two. To determine the appropriate degrees of freedom, we note that the saturated model has IJK parameters, while the model under investigation has $1+(I-1)+(J-1)+(K-1)+(J-1)(K-1)$ parameters, so the degrees of freedom for our test is $IJK-JK-I+1$. Of course, completely analogous procedures are available for testing the independence of any of the variable from the remaining two.

Example 3 (Continued): We can easily test each of the three different independence hypotheses:

```
> aitot <- apply(yijk,c(1,2),sum)
> ahtot <- apply(yijk,c(1,3),sum)
> ihtot <- apply(yijk,c(2,3),sum)
> eijk12 <- array(0,c(2,2,2),dimnames=list(c("<45yo",">45yo"),
+   c("ins.yes","ins.no"),c("hist.yes","hist.no")))
> eijk13 <- array(0,c(2,2,2),dimnames=list(c("<45yo",">45yo"),
+   c("ins.yes","ins.no"),c("hist.yes","hist.no")))
> eijk23 <- array(0,c(2,2,2),dimnames=list(c("<45yo",">45yo"),
+   c("ins.yes","ins.no"),c("hist.yes","hist.no")))
> eijk12[,,1] <- histtot[1]*aitot/tot
> eijk12[,,2] <- histtot[2]*aitot/tot
> eijk13[,1,] <- instot[1]*ahtot/tot
> eijk13[,2,] <- instot[2]*ahtot/tot
> eijk23[1,,] <- agetot[1]*ihtot/tot
> eijk23[2,,] <- agetot[2]*ihtot/tot
> pres12 <- (yijk-eijk12)/sqrt(eijk12)
> pres13 <- (yijk-eijk13)/sqrt(eijk13)
> pres23 <- (yijk-eijk23)/sqrt(eijk23)
> c(sum(pres12^2),1-pchisq(sum(pres12^2),8-4-2+1))
[1] 1.8749436 0.5987637
> c(sum(pres13^2),1-pchisq(sum(pres13^2),8-4-2+1))
[1] 5.230660e+01 2.576683e-11
> c(sum(pres23^2),1-pchisq(sum(pres23^2),8-4-2+1))
[1] 5.367641e+01 1.315292e-11
```

So, it appears that family history is independent of age at onset and insulin dependence (indeed, if we look at the table, we see that for those with a family history, the proportion of early onset diabetics who were insulin dependent was 6/7=85.7% and of late onset diabetics was 6/42=14.3%, while among those without a family history, the proportions of insulin dependents were 16/18=88.9% and 8/56=14.3% for early and late onset diabetics, respectively). However, clearly insulin dependence and age at onset are strongly dependent on each other, and the observed counts clearly indicate that the dependence is in the direction of those with late onset ages being much less likely to be insulin dependent. In fact, since the variables are independent of family history, it makes sense to simply collapse the data over these two categories and analyse the data as if it were a two-way table. Thus, our estimate of the proportion of early onset diabetics who are insulin dependent is 22/25=88% and the proportion of late onset diabetics who are insulin dependent is 14/98=14.3%.

However, we close with an example which shows that collapsing tables (and thus ignoring variables) is dangerous and can lead to very misleading conclusions if the variable being ignored is not independent of the rest of the variables.

Example 4 - Berkeley Graduate Admissions Data: A (fictitious) sample of 420 graduate applications were broken down by the gender of the applicant, the faculty of application and the success of the application:

		Application Result	
Faculty	Gender	Admit	Deny
Science	Male	90	90
	Female	30	30
Arts	Male	15	45
	Female	30	90

The question of interest is whether there is a gender bias in the admissions procedure. An examination of the above data clearly shows that the answer is that no bias exists. Within the science faculty 50% of applications are successful regardless of the gender of the applicant, while in the arts faculty 25% of applicants are successful regardless of gender. However, if we collapse the data over both the faculties we see that:

```
> yijk <- array(0,c(2,2,2),dimnames=list(c("male","female"),
+   c("admit","deny"),c("science","arts")))
> yijk[,,1] <- matrix(c(90,30,90,30),ncol=2)
> yijk[,,2] <- matrix(c(15,30,45,90),ncol=2)
> oij <- apply(yijk,c(1,2),sum)
> oij
       admit deny
  male   105  135
female    60  120
> gtot <- apply(oij,1,sum)
> atot <- apply(oij,2,sum)
> eij <- gtot%*%t(atot)/sum(atot)
> chi2 <- sum(((oij-eij)^2)/eij)
> c(chi2,1-pchisq(chi2,1))
[1] 4.67914439 0.03053095
```

So, it appears from this analysis that there is a relationship between gender and success of applications; namely that 105/240=43.75% of male applicants are successful and only 60/180=33.33% of female applicants are successful. Of course, the problem here is that we have collapsed the tables over a variable which is not independent of the other two:

```
> eijk12 <- array(0,c(2,2,2),dimnames=list(c("male","female"),
+   c("admit","deny"),c("science","arts")))
> ftot <- apply(yijk,3,sum)
> agtot <- apply(yijk,c(1,2),sum)
> eijk12[,,1] <- ftot[1]*agtot/sum(ftot)
> eijk12[,,2] <- ftot[2]*agtot/sum(ftot)
> chi2 <- sum(((yijk-eijk12)^2)/eijk12)
> c(chi2,1-pchisq(chi2,3))
[1] 91.875  0.000
```

Of course, what is happening here is that the arts faculty is the more difficult one to which to gain admission, and most of the female applicants apply to this faculty. Thus, if we ignore the faculty to which each the application was sent, it appears that females are being discriminated against since they tend to apply to the faculty with the lower admissions rate. This phenomenon is known as

Simpson's paradox and is a special case of a more general statistical concept known as *confounding*. A more appropriate analysis of whether gender and admission are related in this case would be to test the hypothesis of "conditional independence", which has the form $H_0 : \pi_{ijk} = \pi_{i\bullet k}\pi_{\bullet jk}/\pi_{\bullet\bullet k}$. Under this null hypothesis, the expected cell counts would be:

$$E_{ijk} = n\left(\frac{Y_{i\bullet k}}{n}\right)\left(\frac{Y_{\bullet jk}}{n}\right)\left(\frac{n}{Y_{\bullet\bullet k}}\right) = \frac{Y_{i\bullet k}Y_{\bullet jk}}{Y_{\bullet\bullet k}},$$

new expected cell counts

and thus the appropriate Pearson's chi-squared analysis is:

```
> yi.k <- apply(yijk,c(1,3),sum)
> y.jk <- apply(yijk,c(2,3),sum)
> y..k <- apply(yijk,3,sum)
> eijk <- array(0,c(2,2,2),dimnames=list(c("male","female"),
+   c("admit","deny"),c("science","arts")))
> eijk[,,1] <- yi.k[,1]%*%t(y.jk[,1])/y..k[1]
> eijk[,,2] <- yi.k[,2]%*%t(y.jk[,2])/y..k[2]
> chi2 <- sum((yijk-eijk)^2/eijk)
> c(chi2,1-pchisq(chi2,1))
[1] 0 1
```

And we see that the null hypothesis is now not rejected (indeed, the Pearson's chi-squared statistic is exactly zero in this case!), so within each faculty (which is the idea of a "conditional" statement) the gender and admission variables are independent. As a final note, we see that the 10 probabilities $\pi_{1\bullet 1}$, $\pi_{1\bullet 2}$, $\pi_{2\bullet 1}$, $\pi_{2\bullet 2}$, $\pi_{\bullet 11}$, $\pi_{\bullet 12}$, $\pi_{\bullet 21}$, $\pi_{\bullet 22}$, $\pi_{\bullet\bullet 1}$, $\pi_{\bullet\bullet 2}$ satisfy the four relationships

$$\pi_{1\bullet 1} + \pi_{2\bullet 1} = \pi_{\bullet\bullet 1}; \qquad \pi_{\bullet 11} + \pi_{\bullet 21} = \pi_{\bullet\bullet 1};$$
$$1 - \pi_{1\bullet 1} - \pi_{2\bullet 1} - \pi_{1\bullet 2} = \pi_{2\bullet 2}; \qquad 1 - \pi_{\bullet 11} - \pi_{\bullet 12} - \pi_{\bullet 21} = \pi_{\bullet 22},$$

implying that we really only need to know 6 "effective" parameters. Thus, the appropriate degrees of freedom for the test of "conditional independence" are $7 - 6 = 1$, since the full model has 8 π_{ijk}'s leading to 7 "effective" parameters, since all the π_{ijk}'s must sum to unity.