Three-way Contingency Tables

Statistics 149

Spring 2006

Three-way Contingency Tables

Example: Classroom Behaviour (Everitt, 1977, page 67)

97 students were classified on three factors

- X: Teacher's rating of classroom behaviour (`behaviour`) - non deviant or deviant
- Y: Risk index based on home conditions (`risk`) - not at risk or at risk
- Z: Adversity of school conditions (`adversity`) - low, medium, or high

Adversity	Low		Medium		High	
Risk	Not at Risk	At Risk	Not at Risk	At Risk	Not at Risk	At Risk
Non Deviant	16	7	15	34	5	3
Deviant	1	1	3	8	1	3

Mimicing the earlier notation, the cell counts n_{ijk} are the number observations where $X = x_i, Y = y_j, Z = z_k$ where $i = 1, \ldots, I; j = 1, \ldots, J; k = 1, \ldots, K$.

The question of interest is what is the relationship between X, Y, and Z. In the example, the relationship between `behaviour`, `risk`, and `adversity`.

To examine this, we will examine the set of hierarchical log linear models on π_{ijk}, where

$$\pi_{ijk} = P[X = x_i, Y = y_j, Z = z_k]; \qquad \mu_{ijk} = n\pi_{ijk}$$

Remember in hierarchical models, if an interaction is contained in the model, all lower order interactions and main effects must also be contained in the model.

In the case of only two factors, there are only two models of usual interest, the saturated and independence (homogeneity) models.

When there are three or more factors, the classes of models in much more interesting.

In what follows, one notation that is used to describe a model is based on the highest order interactions in the model such that all terms in the model are implied. For example

$$(XY, XZ)$$

describes the model with the XY and XZ interactions, and the X, Y, and Z main effects. This notation relates to compact forms for writing the models in **R**. This model in **R** could be written as

```
n ~ X*Y + X*Z
```

Saturated: (XYZ)

$$\log \pi_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ}$$

In this case there is no nice relationship between the three variables. All the information about cell i, j, k is given by the count n_{ijk}.

The fits for this model satisfy

$$\hat{\pi}_{ijk} = \frac{n_{ijk}}{n}; \qquad \hat{\mu}_{ijk} = n_{ijk}$$

The number of parameters to be fit are

Parameters	# terms
λ	1
λ_i^X	$I-1$
λ_j^Y	$J-1$
λ_k^Z	$K-1$
λ_{ij}^{XY}	$(I-1)(J-1)$
λ_{ik}^{XZ}	$(I-1)(J-1)$
λ_{jk}^{YZ}	$(J-1)(K-1)$
λ_{ijk}^{XYZ}	$(I-1)(J-1)(K-1)$
Total	IJK

As with the two-way models, there are constraints that need to be considered when parameterizing the λs, which is where the -1 terms come in the above table.

As the total number of parameters in the model = the number of cells in the table, the degrees of freedom for this model is 0.

All the other models to be considered are subsets of this model where different combinations of the λs are set to 0.

Homogeneous association: (XY, XZ, YZ)

This model is derived by setting all $\lambda_{ijk}^{XYZ} = 0$, giving

$$\log \pi_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$$

For this model, the MLE equations imply

$$\hat{\mu}_{ij+} = n_{ij+} \qquad \hat{\mu}_{i+k} = n_{i+k} \qquad \hat{\mu}_{+jk} = n_{+jk}$$

So we can figure out what happends when we collapse over one variable.

However there are no direct estimates for $\hat{\mu}_{ijk}$. To get these you need some sort of iterative scheme, such as Newton-Raphson, iteratively reweighted least-squares, or iterative proportional fitting.

Interpretation of this model: Since the λ_{ijk}^{XYZ} in the saturated model measures the difference between 2-factor effects attributable to a third variable, setting all $\lambda_{ijk}^{XYZ} = 0$ describes a table with constant 2-factor effects, though not necessarily 0.

Can think of this as a partial association model.

Suppose the model holds. Lets fix a level $Z = z_k$ and look at an odds ratio involving X and Y.

$$
\begin{aligned}
\phi_{(ii')(jj')|k} &= \log \frac{\mu_{ijk}\mu_{i'j'k}}{\mu_{i'jk}\mu_{ij'k}} = \log \frac{\pi_{ijk}\pi_{i'j'k}}{\pi_{i'jk}\pi_{ij'k}} \\
&= \quad \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} \\
&\quad + \lambda + \lambda_{i'}^X + \lambda_{j'}^Y + \lambda_k^Z + \lambda_{i'j'}^{XY} + \lambda_{i'k}^{XZ} + \lambda_{j'k}^{YZ} \\
&\quad - \lambda - \lambda_{i'}^X - \lambda_j^Y - \lambda_k^Z - \lambda_{i'j}^{XY} - \lambda_{i'k}^{XZ} - \lambda_{jk}^{YZ} \\
&\quad - \lambda - \lambda_i^X - \lambda_{j'}^Y - \lambda_k^Z - \lambda_{ij'}^{XY} - \lambda_{ik}^{XZ} - \lambda_{j'k}^{YZ} \\
&= \lambda_{ij}^{XY} - \lambda_{i'j}^{XY} - \lambda_{ij'}^{XY} + \lambda_{i'j'}^{XY}
\end{aligned}
$$

First note that this does not depend on k, just on i, i', j, j'.

There are three sets of these partial association measures for a three-way table, one for each variable conditioned on.

For the school behaviour example, lets examine the relationship between `behaviour` and `risk` for each level of `adversity`.

Adversity	Low		Medium		High	
Risk	Not at Risk	At Risk	Not at Risk	At Risk	Not at Risk	At Risk
Non Deviant	16	7	15	34	5	3
Deviant	1	1	3	8	1	3

$$\hat{\phi}_{\text{Low}} = \frac{16 \times 1}{1 \times 7} = 2.29; \qquad \hat{\phi}_{\text{Med}} = \frac{15 \times 8}{3 \times 34} = 1.18; \qquad \hat{\phi}_{\text{High}} = \frac{5 \times 3}{1 \times 3} = 5$$

While their appears to be some difference between these, remember that the sample sizes are small.

Adversity	Low	Medium	High
$\hat{\phi}$	0.83	0.16	1.61
$\mathrm{SE}(\hat{\phi})$	1.24	0.70	1.19

For this model $df = (I-1)(J-1)(K-1)$, which happens to be the number of λs set to zero in the saturated log linear model.

Conditional independence: (XY, XZ), (XY, YZ), or (XZ, YZ)

There are three different models of this form, which can be derived by dropping the three-factor interactions and one set of two-factor interactions.

One of the possible forms (for the model (XY, YZ)) is

$$\log \pi_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ}$$

in this case the λ_{ijk}^{XYZ}s and the λ_{ik}^{XZ}s are all set to zero.

The degrees of freedom for this model is $(I-1)(K-1)J$. Again this is the number of λs set to zero.

Models of this form correspond to conditional independence of the variables. For the example (XY, YZ), X and Z are conditionally independent given the level of Y.

It can be shown that

$$P[X = x_i, Z = z_k | Y = y_j] = P[X = x_i | Y = y_j] P[Z = z_k | Y = y_j]$$

for each level of Y under this model.

One way of thinking of this, if you look at the J different $I \times K$ tables you get by fixing the level of Y and classifying by X and Z, each of them exhibits independence.

So one consequence of this is that

$$P[X = x_i, Y = y_j, Z = z_k] = P[X = x_i | Y = y_j] P[Z = z_k | Y = y_j] P[Y = y_j]$$

(actually showing that this relationship holds proves the conditional independence assumption)

So given the labeling of X, Y, and Z, (XY, YZ) corresponds to `behaviour` being independent of `adversity` given the risk status of the child.

Note that this is not equivalent to `behaviour` and `adversity` being independent. This form of independence fails as $P[X = x_i|Y = y_j]$ and $P[Z = z_k|Y = y_j]$ could have different forms for each level y_j.

If these vary

$$
\begin{aligned}
P[X = x_i, Z = z_k] &= \sum_{j=1}^{J} P[X = x_i|Y = y_j]P[Z = z_k|Y = y_j]P[Y = y_j] \\
&\neq P[X = x_i]P[Z = z_k]
\end{aligned}
$$

This model does have explicit solutions. First the MLE conditions imply that

$$\hat{\mu}_{ij+} = n_{ij+}; \qquad \hat{\mu}_{+jk} = n_{+jk}$$

Next it can be shown that

$$\hat{\mu}_{ijk} = \frac{n_{ij+}n_{+jk}}{n_{+j+}}$$

What's the difference between (XY, XZ, YZ) and (XZ, YZ)?

In (XY, XZ, YZ), the values

$$\phi_{(ii')(jj')|k} = c_{(ii')(jj')}$$

take the same value for each k. (They will change when i, i', j, j' change)

For the model (XZ, YZ),

$$\phi_{(ii')(jj')|k} = 1$$

for all k.

This model corresponds to the hypothesis being testing by the Mantel-Haenszel test.

Partial independence: (XY, Z), (XZ, Y), or (YZ, X)

The next model along the line comes from dropping the three-factor interaction and two of the two-factor interactions. Another way of thinking of it is that the model only includes one two-factor interaction and the main effects of all variables.

Again there are three different models of this form, one for each of the possible two-way interactions. The general form of the model is

$$\log \pi_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{jk}^{YZ}$$

(this is for (X, YZ)).

This model has $df = (I-1)(JK-1)$

In these models, one variable is independent of the combination of the other two variables. So (X, YZ) corresponds to `behaviour` being independent of the `risk`, `adversity` combinations.

So under this model

$$P[X = x_i, Y = y_j, Z = Z_k] = P[X = x_i]P[Y = y_j, Z = Z_k]$$

In this model Y and Z are dependent, even after marginalizing out X. However X and Y are independent (similarly for X and Z). This holds since

$$\begin{aligned} P[X = x_i, Y = y_j] &= \sum_{k=1}^{K} P[X = x_i, Y = y_j, Z = Z_k] \\ &= \sum_{k=1}^{K} P[X = x_i]P[Y = y_j, Z = Z_k] \\ &= P[X = x_i] \sum_{k=1}^{K} P[Y = y_j, Z = Z_k] \\ &= P[X = x_i]P[Y = y_j] \end{aligned}$$

For this model the following relationships between the fits can be shown

$$\hat{\mu}_{i++} = n_{i++} \qquad \hat{\mu}_{+jk} = n_{+jk}$$

and

$$\hat{\mu}_{ijk} = \frac{n_{i++}n_{+jk}}{n_{+++}}$$

Mutual independence: (X, Y, Z)

This model only includes the three sets of main effects. The form of the log linear model is

$$\log \pi_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z$$

For this model $df = IJK - \{(I-1) + (J-1) + (K-1) + 1\} = IJK - I - J - K - 2$.

Under this model, there is mutual independence of all three variables. From this it can be shown that

$$\hat{\mu}_{i++} = n_{i++} \qquad \hat{\mu}_{+j+} = n_{+j+} \qquad \hat{\mu}_{++k} = n_{++k}$$

and

$$\hat{\mu}_{ijk} = \frac{n_{i++} n_{+j+} n_{++k}}{(n_{+++})^2}$$

Since all three variables are mutually independent, it follows that any pair must be as well.

Noncomprehensive

These models have the feature that at least one variable is not included. Two examples are (XY) and (X, Y).

One way of thinking of these models is collapsing the table over the dropped variable. So (XY) and (X, Y) involve collapsing over Z, i.e. look at the two-way table with counts n_{ij+}.

In both cases

$$\mu_{ijk} = \frac{\mu_{ij+}}{K}$$

However the form of μ_{ij+} will vary depending on the form of association between X and Y.

Comparing Fits in Example

Model	df	X_p^2	X^2	p-value(X^2)	
(BRA)	0	0	0	1	
(BR, BA, RA)	2	0.93	0.94	0.624	✔
(BR, BA)	4	11.17	11.32	0.023	
(BR, RA)	4	4.51	4.12	0.390	✔
(BA, RA)	3	1.87	1.90	0.593	✔
(BR, A)	6	15.07	14.98	0.020	
(BA, R)	5	12.64	12.76	0.026	
(RA, B)	5	6.19	5.56	0.351	✔
(B, R, A)	7	17.30	16.42	0.022	

So there appear to be 4 models that seem to fit adequately.

The four models correspond to

- (BR, BA, RA): conditional odds ratios are the same. In particular the relationship between `behaviour` and `risk` is the same for each level of `adversity`. Similarly the relationship between `behaviour` and `adversity` is the same for each `risk` level

- (BR, RA): given `risk` level, `behaviour` and `adversity` are independent

- (BA, RA): given `adversity` level, `behaviour` and `risk` are independent

- (RA, B): `behaviour` is independent of `risk` and `adversity`

However do we get a significantly worse fit with the smaller models? Lets compare the nested models by the drop in deviance test.

```
> anova(behave.xy.xz.yz, behave.xy.yz, test='Chisq')
Analysis of Deviance Table

Model 1: n ~ (behaviour + risk + adversity)^2
Model 2: n ~ behaviour * risk + risk * adversity
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1         2     0.9428
2         4     4.1180 -2  -3.1752    0.2044

> anova(behave.xy.xz.yz, behave.xz.yz, test='Chisq')
Analysis of Deviance Table

Model 1: n ~ (behaviour + risk + adversity)^2
Model 2: n ~ behaviour * adversity + risk * adversity
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1         2    0.94285
2         3    1.90396 -1 -0.96112   0.32691
```

These two tests imply we don’t need all three interactions

```
> anova(behave.xy.xz.yz, behave.yz.x, test='Chisq')
Analysis of Deviance Table

Model 1: n ~ (behaviour + risk + adversity)^2
Model 2: n ~ risk * adversity + behaviour
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1          2     0.9428
2          5     5.5603 -3  -4.6175    0.2020

> anova(behave.xy.yz, behave.yz.x, test='Chisq')
Analysis of Deviance Table

Model 1: n ~ behaviour * risk + risk * adversity
Model 2: n ~ risk * adversity + behaviour
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1          4     4.1180
2          5     5.5603 -1  -1.4423    0.2298
```

```
> anova(behave.xz.yz, behave.yz.x, test='Chisq')
Analysis of Deviance Table

Model 1: n ~ behaviour * adversity + risk * adversity
Model 2: n ~ risk * adversity + behaviour
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1         3     1.9040
2         5     5.5603 -2  -3.6563    0.1607
```

These three tests imply that the partial independence seems reasonable.

The reason for doing the model comparisons here in addition to the goodness of fit tests is that drop in deviance tests tend to give better information about the significance of fits.

It is possible to have two nested models, both with insignficant goodness of fit tests, but the drop in deviance test suggest that the smaller model is not adequate for describing the situation.

Note that the mutual independence model does not fit, implying association between `risk` and `adversity`.

Let collapse across `behaviour`, giving the table

	Low	Medium	High	Total
Not at Risk	17	18	6	41
At Risk	8	42	6	56
Total	25	60	12	97

```
> anova(behave.y.z, behave.yz, test='Chisq')
Analysis of Deviance Table

Model 1: n ~ risk + adversity
Model 2: n ~ risk * adversity
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1         8     60.849
2         6     49.990  2   10.859     0.004
```

This implies that there is an association between `risk` and `adversity`.

Note that the test done above is equivalent to the test

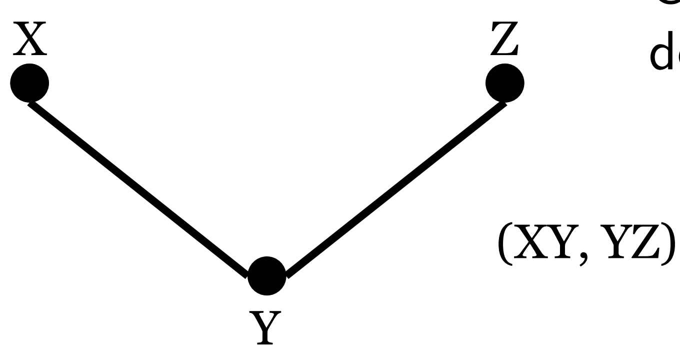
```
> anova(behave.x.y.z, behave.yz.x, test='Chisq')
Analysis of Deviance Table

Model 1: n ~ behaviour + risk + adversity
Model 2: n ~ risk * adversity + behaviour
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1         7    16.4192
2         5     5.5603  2  10.8589    0.0044
```

This is an example showing that it ok to collapse across a variable that is independent of the rest.

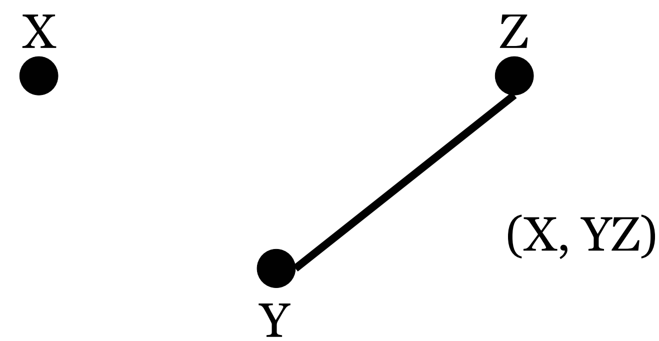
Describing Independence Relationships Graphically

There is an approach to describing the independence relationships in a model based on graph theory. The set of hierarchical log linear models for describing contingency tables are examples of graphical models, whose probability structure can be described by a graph.

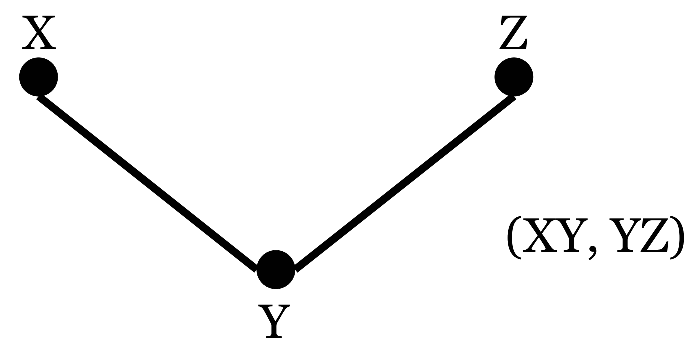
The idea is that the factors in a model are used to give the nodes of a graph and interaction terms in the model give the edges.



Consider the model (XY, YZ). This can be described by the graph, Similarly the model (X, YZ) can be described by the graph,

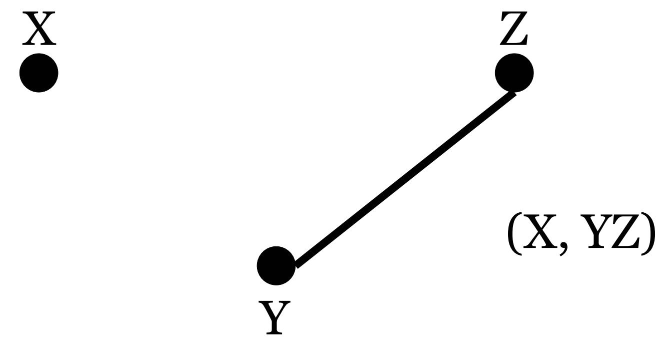


To describe determine is a pair of variables are independent, you need to see if there is a path in the graph joining the variables. If there is, the variables are not independent. If not, independent



So based on the graph just to the above right, X is not independent of Y or Z.

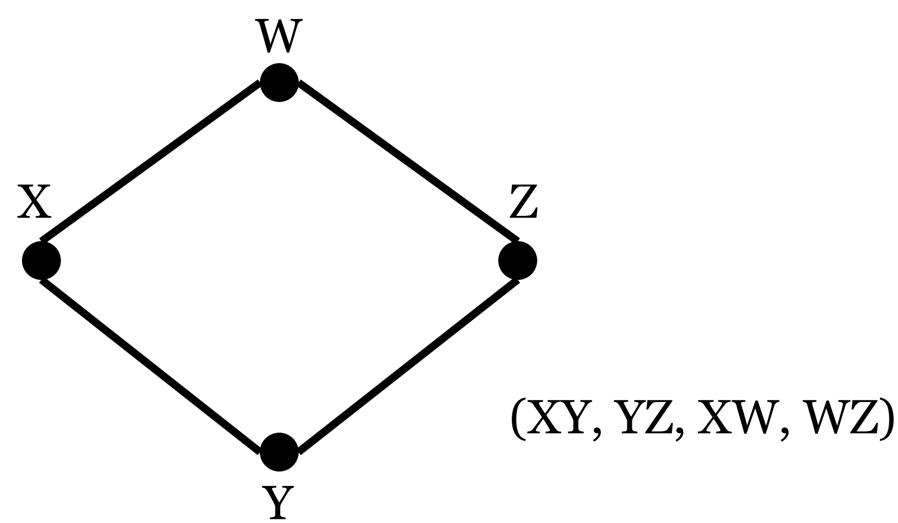
However for this model, we can see that X is independent of Y and Z, but that Y and Z are associated.



Conditional independence is examined by looking to see if a set of nodes separate two other nodes (or two other sets of nodes).

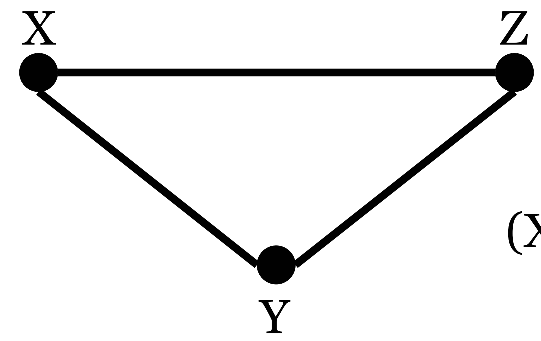
In this graph, the nodes for X and Z are separated by the nodes Y and W. So this graph corresponds to X and Z being conditionally independent, given Y and W.

However X is not conditionally independent of Z given only Y since I can go from X to Z without having to go through the conditioning set.



Note that graphs are not unique. There are cases where different models give the same graphs.

For example the models (XY, XZ, YZ) and (XYZ) are both described by the graph.



Note that these graphical ideas are quite general and can be used in many situations. For example, in large Bayesian models, the graphical structures can be used to help design Gibbs samplers, as they immediately show what variables need to be conditioned on in each step.

Getting back to log-linear models on contingency tables, these graphs can be used to help determine whether closed form solutions for $\hat{\mu}$ exist. As models get more complicated, often the $\hat{\mu}$ need to be determined by iterative scheme, not by nice formulas like (for the model (X, YZ))

$$\hat{\mu}_{ijk} = \frac{n_{i++}n_{+jk}}{n_{+++}}$$