R Tutorial: Partition LR Chi-Square I x J (5 x 2) Contingency Table

W. Gregory Alvord
National Cancer Institute

We partition (decompose) a 5 x 2 contingency table with R. The data are from Alan Agresti's ***Categorical Data Analysis*** (1990, 2nd ed.) – Table 3.10, Problem 3.6, page 72.

```
> Ag.3.10.table.entries <- c(105,12,18,47,0,8,2,19,52,13)
```

```
> Ag.3.10.table <- as.table(matrix(Ag.3.10.table.entries,
nrow = 5, byrow = FALSE, dimnames = list(Diagnosis =
c('Schizophrenia', 'Affective.Disorder', 'Neurosis',
'Personality.Disorder', 'Special.Symptoms'), Drugs.Rx =
c('Yes', 'No'))))
```

```
> addmargins(Ag.3.10.table) # add marginal sums to table
                      Drugs.Rx
Diagnosis              Yes  No Sum
  Schizophrenia        105   8 113
  Affective.Disorder    12   2  14
  Neurosis              18  19  37
  Personality.Disorder  47  52  99
  Special.Symptoms       0  13  13
  Sum                  182  94 276
```

Two hundred seventy-six (276) psychiatric patients were cross classified as to their diagnosis in one of five psychiatric groups: (1) Schizophrenia, (2) Affective Disorder, (3) Neurosis, (4) Personality Disorder, and (5) Special Symptoms and as to whether (or not) they were prescribed drugs in their treatment regimens.

We examine the *relationship* between the patients' diagnostic class (**`Diagnosis`**) and whether or not drugs were prescribed (**`Drugs.Rx`**). We use the Likelihood Ratio Chi-Squared statistic (as opposed to the Pearson statistic), also known as LR χ^2 (LR X^2, G^2) to test for independence between **`Diagnosis`** and **`Drugs.Rx`**. (Independent partitionings of χ^2 have the property that their LR values and degrees of freedom are additive (Agresti, 1990, pp 50-51)). One way to do this is with the **`loglm()`** function from the **`MASS`** package. First, load the MASS package (Venables and Ripley). Next, perform a 'global' test for the hypothesis of independence (no association) between **`Diagnosis`** and **`Drugs.Rx`**. The null hypothesis states that **`Diagnosis`** and **`Drugs.Rx`** are independent.

```
> library(MASS) # Venables and Ripley
```

```
> Ag.3.10.global.loglm <- loglm( ~ Diagnosis + Drugs.Rx,
 data = Ag.3.10.table)
```

```
> Ag.3.10.global.loglm
```

```
Call:
loglm(formula = ~Diagnosis + Drugs.Rx, data = Ag.3.10.table)

Statistics:
                      X^2 df P(> X^2)
Likelihood Ratio 96.53689  4        0 # p << 0.0001
Pearson          84.18847  4        0
```

Reject the null hypothesis that **Diagnosis** and **Drugs.Rx** are independent; the LR X^2 value is 96.54 on 4 df, p << 0.0001. [Note: Ignore the Pearson X^2 value in these analyses.]

The loglinear analysis reveals a strong relationship between **Diagnosis** and **Drugs.Rx**. However, we wish to ascertain more specifically which diagnostic categories, or groupings of diagnostic categories, account for the relationship. We partition (decompose) the table in a statistically rigorous way to "describe similarities and differences among the diagnoses in terms of the relative frequencies of the prescribed drugs," (Agresti, page 72). The decomposition involves the partitioning of the contingency table and its corresponding Likelihood Ratio Chi-Square statistic, LR χ^2, into independent (orthogonal), additive components (Agresti, pp 50-54). The advantage to this is that independent inferences can be drawn for each component involved in the partitioning. "A [correct] partitioning may show that an association primarily reflects differences between certain categories or groupings of categories," (Agresti, page 50). Rules for partitioning the table are provided in Agresti (page 53).

Search for sub-tables that might be homogeneous, which can be combined (collapsed). For example, identify two rows of this table (i.e., two psychiatric diagnostic groups) that appear to have comparable proportions (percentages) of cases classified as **Yes** (or alternatively as **No**). Prepare a table comprised of percentages that, for each diagnostic group, sum to 100% across the two categories of whether or not drugs were prescribed (**Yes** or **No**).

```
# compute proportions across columns
> Ag.3.10.prop.mar.1.table = prop.table(Ag.3.10.table,
 margin = 1)

# transform to percentages
> Ag.3.10.percent.mar.1.table = 100*Ag.3.10.prop.mar.1.table

# present 'percentages' table
> round(Ag.3.10.percent.mar.1.table, 1) # Percentage
                       Drugs.Rx
Diagnosis               Yes   No
  Schizophrenia        92.9  7.1
  Affective.Disorder   85.7 14.3
  Neurosis             48.6 51.4
  Personality.Disorder 47.5 52.5
  Special.Symptoms       0.0 100.0
```

For **Neurosis**, 48.6% of patients were prescribed drugs (**Yes**) while 51.4% were not (**No**). For **Personality.Disorder**, 47.5% were prescribed drugs while 52.5% were not. The percentages of patients who were prescribed drugs for **Neurosis** (48.6%) and **Personality.Disorder** (47.5%) appear to be comparable. From the original 5 x 2 table of observed frequencies, extract the following 2 x 2 sub-table.

```
> Ag.3.10.rows.34.table <- as.table(Ag.3.10.table[3:4,])

> Ag.3.10.rows.34.table
                      Drugs.Rx
Diagnosis              Yes No
  Neurosis              18 19
  Personality.Disorder  47 52
```

Test for independence between **Diagnosis** and **Drugs.Rx** for these two diagnostic classes alone.

```
> Ag.3.10.rows.34.loglm <- loglm( ~ Diagnosis + Drugs.Rx,
data = Ag.3.10.rows.34.table)

> Ag.3.10.rows.34.loglm
Call:
loglm(formula = ~Diagnosis + Drugs.Rx, data =
Ag.3.10.rows.34.table)

Statistics:
                        X^2 df  P(> X^2)
Likelihood Ratio 0.01487122  1 0.9029405 # ← p = 0.90 n.s.
. . .
```

Do not reject the null hypothesis for independence, LR X^2 = 0.015 on 1 df, p = 0.90.

Now continue the search for other homogeneous patterns in the original 5 x 2 table. Extract the 2 x 2 sub-table considering only those cases associated with **Schizophrenia** and **Affective.Disorder** and test for independence.

```
## Create rows 1 & 2 subtable
> Ag.3.10.rows.12.table <- as.table(Ag.3.10.table[1:2,])

> Ag.3.10.rows.12.table # Observed
                    Drugs.Rx
Diagnosis            Yes  No
  Schizophrenia      105   8
  Affective.Disorder  12   2

## Test for independence in this 2 x 2 sub-table
```

```
> Ag.3.10.rows.12.loglm <- loglm( ~ Diagnosis + Drugs.Rx,
 data = Ag.3.10.rows.12.table)

> Ag.3.10.rows.12.loglm
Call:
loglm(formula = ~Diagnosis + Drugs.Rx, data =
Ag.3.10.rows.12.table)

Statistics:
                       X^2 df  P(> X^2)
Likelihood Ratio 0.7529516  1 0.3855433 ← p = 0.39 n.s.
```

LR X^2 = 0.75 on 1 df, p = 0.39. Do not reject the null.

We have identified two 2 x 2 sub-tables from the original that are homogeneous, one for **`Neurosis`** and **`Personality.Disorder`** and one for **`Schizophrenia`** and **`Affective.Disorder`**. When this occurs the counts in the sub-table can be combined or 'collapsed', i.e., summed over its margins, without loss of information. The original 5 x 2 table can now be collapsed (combined) into a 3 x 2 table.

```
> apply(Ag.3.10.table[1:2,], 2, sum)
> apply(Ag.3.10.table[3:4,], 2, sum)

> Ag.3.10.collapsed.entries <- c(117, 65, 0, 10, 71, 13)

> Ag.3.10.collapsed.mat <- matrix(Ag.3.10.collapsed.entries,
 nrow = 3, byrow = FALSE, dimnames = list(Diagnosis =
 c('Schiz.or.Aff.Dis', 'Neur.or.Pers.Dis',
 'Special.Symptoms'), Drugs.Rx = c('Yes', 'No')))

> Ag.3.10.collapsed.table <- as.table(Ag.3.10.collapsed.mat)

> Ag.3.10.collapsed.table # Observed
                  Drugs.Rx
Diagnosis            Yes  No
  Schiz.or.Aff.Dis 117  10 # combining Schiz & Aff.Dis
  Neur.or.Pers.Dis  65  71 # combining Neur & Pers.Dis
  Special.Symptoms   0  13 # original Spec Sympt counts
```

The observed counts for **`Schizophrenia`** and **`Affective.Disorder`** are combined into a single category now labeled **`Schiz.or.Aff.Dis`**. Similarly, the observed counts for **`Neurosis`** and **`Personality.Disorder`** are combined into a single category now labeled **`Neur.or.Pers.Dis`**. Since the counts associated with **`Special.Symptoms`** have not been used in a previous sub-table, they are retained in the table here.

Test for independence between **Diagnosis** and **Drugs.Rx** in the combined table. Actually, we are less concerned with independence here; we compute the LR X^2 statistic to complete the steps for the partitioning.

```
> Ag.3.10.collapsed.loglm <- loglm( ~ Diagnosis + Drugs.Rx,
 data = Ag.3.10.collapsed.table)

> Ag.3.10.collapsed.loglm
Call:
loglm(formula = ~Diagnosis + Drugs.Rx, data =
Ag.3.10.collapsed.table)

Statistics:
                      X^2 df P(> X^2)
Likelihood Ratio 95.76907  2        0 🡐 p << 0.0001
```

Reject the null hypothesis for independence, LR X^2 = 95.77 on 2 df, p << 0.0001.

Summarizing to this point: (1) with respect to the original table, **Diagnosis** and **Drugs.Rx** are not independent (not homogeneous); (2) patients diagnosed with either Schizophrenia or Affective Disorder are homogeneous; (3) patients diagnosed with Neurosis or Personality Disorder are homogeneous; (4) from the combined (collapsed) table, **Diagnosis** and **Drugs.Rx** are not homogeneous.

When the partitioning is performed correctly, the LR X^2 values of the sub-tables sum, exactly, to the LR X^2 value for the original table (Agresti, 1990, pp. 50-51). Similarly, the degrees of freedom associated with each test sum to the degrees of freedom associated with the test from the original table.

Add the three LR X^2 values associated with the three sub-tables . . .

```
> Ag.3.10.rows.12.loglm$lr + Ag.3.10.rows.34.loglm$lr +
Ag.3.10.collapsed.loglm$lr
[1] 96.53689
```

... and compare to LR X^2 value for the original 5 x 2 table, on 4 degrees of freedom,

```
> Ag.3.10.global.loglm$lr
[1] 96.53689

# check accuracy to 12 decimal places

> round(Ag.3.10.rows.12.loglm$lr + Ag.3.10.rows.34.loglm$lr
+ Ag.3.10.collapsed.loglm$lr, 12) ==
round(Ag.3.10.global.loglm$lr, 12)
[1] TRUE
```

They are equal. Also, the degrees of freedom for each component are, respectively, 1, 1, and 2, which sum to 4 degrees of freedom associated with the original table.

Summary and Interpretation

Psychiatric patients were relatively more or less likely to be prescribed drugs depending on their respective diagnoses. Patients diagnosed with Schizophrenia or Affective Disorder were *more* likely to be prescribed drugs than not (92% vs. 8%). Patients diagnosed with Neurosis or Personality Disorder, were about *equally* likely to be prescribed drugs or not (48% vs. 52%). Patients with Special Symptoms were *not* likely to be prescribed drugs; in fact, no drugs were prescribed for these patients in this sample (0.0% vs. 100.0%).

References

Agresti A, *Categorical Data Analysis, 2nd ed.*, Wiley, New York, 1990.

R Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/

Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0

Alvord WG, "*Partition (Decompose) a 5 x 2 Contingency Table using R*" provides a more thorough (and verbose) analysis of the problem considered here: http://css.ncifcrf.gov/services/alvord/PartitionDecompose5x2ContingencyTableWithR.pdf