36-720: Introduction

Brian Junker

August 27, 2007

- Discrete Multivariate Analysis
- Contingency Tables
- 2×2 Tables
- $I \times J$ Tables
- Sampling Models
- Some Sampling Theory

Discrete Multivariate Analysis

- Basic data is discrete: counts, ordinal responses, nominal responses
  - Poisson counts of some event
  - Cross-classified tables of counts (contingency tables)
  - Success/failure records
  - Questionnaire responses
  - Judge’s ratings
  - Classification of objects

Two Basic Kinds of Problems

- Describe/understand structure of a (discrete) multivariate distribution
  - Describe or display associations between variables (log-linear models, graphical models)
  - Find underlying (latent) structure (latent variable models)
- “Generalizations” of regression, with predictors and response variables
  - Response variables are discrete
  - Predictor variables can be
    - Discrete (log-linear models)
    - Mixture of discrete and continuous (logistic regression)

Contingency Tables

- Observed counts n_{ij}, sometimes x_{ij}
- Expected counts m_{ij}, sometimes μ_{ij}
- Cell probabilities (joint distribution) p_{ij}, sometimes π_{ij}
- Marginal probabilities (distribution) p_{i+}, p_{+j} or $p_{i\cdot}$, $p_{\cdot j}$
- Marginal (observed or expected) counts, e.g. n_{i+}, n_{+j} or $n_{i\cdot}$, $n_{\cdot j}$
- Conditional probabilities (distribution) $p_{i|j}$, $\pi_{j|i}$
- Observed probabilities or proportions: $\hat{p}_{ij}$ or $\hat{\pi}_{ij}$, etc.

OBSERVED COUNTS

	Myocardial Infarction[*]			
	Fatal	Nonfatal	No	
	Attack	Attack	Attack	Total
Placebo	$n_{11} = 18$	$n_{12} = 171$	$n_{13} = 10845$	$n_{1+} = 11034$
Aspirin	$n_{21} = 5$	$n_{22} = 99$	$n_{23} = 10993$	$n_{2+} = 11097$
Total	$n_{+1} = 23$	$n_{+2} = 270$	$n_{+3} = 21838$	$n_{++} = 22131$

[*] Physicians’ Health Study Research Group, Harvard Medical School.

EXPECTED COUNTS

	Myocardial Infarction			
	Fatal	Nonfatal	No	
	Attack	Attack	Attack	Total
Placebo	m_{11}	m_{12}	m_{13}	m_{1+}
Aspirin	m_{21}	m_{22}	m_{23}	m_{2+}
Total	m_{+1}	m_{+2}	m_{+3}	m_{++}

PROBABILITIES

	Myocardial Infarction			
	Fatal	Nonfatal	No	
	Attack	Attack	Attack	Total
Placebo	p_{11}	p_{12}	p_{13}	p_{1+}
Aspirin	p_{21}	p_{22}	p_{23}	p_{2+}
Total	p_{+1}	p_{+2}	p_{+3}	p_{++}

- For a *designed experiment* we might have independent fixed rows with fixed totals n_{i+}. Then it makes sense to have $p_{ij} = P[j|i]$ and hence $m_{ij} = n_{i+}p_{ij}$ (so $p_{i+} = 1$, and neither p_{+j} nor p_{++} have useful interpretations);
- For an *observational study* we may have only the grand total n_{++} fixed. Then it makes sense to have $p_{ij} = P[i \,\&\, j]$ and hence $m_{ij} = n_{++}p_{ij}$, and $p_{++} = 1$.

2×2 **Tables**

Let us assume that the subtable

	Myocardial Infarction		
	Fatal Attack	Nonfatal Attack	Total
Placebo	$n_{11} = 18$	$n_{12} = 171$	$n_{1+} = 189$
Aspirin	$n_{21} = 5$	$n_{22} = 99$	$n_{2+} = 104$
Total	$n_{+1} = 23$	$n_{+2} = 270$	$n_{++} = 293$

represents a *designed experiment* with *independent Binomial rows*.

Is the treatment effective? We want to test

$$H_0 : \; p_{11} = p_{21} \text{ and } p_{12} = p_{22}$$

Since the rows are binomial we only need to test $H_0 : \; p_{11} - p_{21} = 0$.

Two-Sample Test

$\hat{p}_{11} - \hat{p}_{21} = \dfrac{n_{11}}{n_{1+}} - \dfrac{n_{21}}{n_{2+}}; \qquad E[\hat{p}_{11} - \hat{p}_{21}] = p_{11} - p_{21}$

$\text{Var}\,(\hat{p}_{11} - \hat{p}_{21}) = \dfrac{p_{11}(1 - p_{11})}{n_{1+}} + \dfrac{p_{21}(1 - p_{21})}{n_{2+}};$

If $H_0 : \; p_{11} = p_{21} = p$ holds, then $\text{Var}\,(\hat{p}_{11} - \hat{p}_{21}) = p(1-p) \cdot \left[\dfrac{1}{n_{1+}} + \dfrac{1}{n_{2+}}\right];$

$z = \dfrac{\hat{p}_{11} - \hat{p}_{21}}{\sqrt{\overline{p}(1 - \overline{p})\left[\frac{1}{n_{1+}} + \frac{1}{n_{2+}}\right]}} \overset{approx}{\sim} N(0, 1)$, where $\overline{p} = \dfrac{n_{11} + n_{21}}{n_{1+} + n_{2+}}$.

In our case, $\hat{p}_{11} = n_{11}/n_{1+} = 18/189 = 0.095$,
$\hat{p}_{21} = n_{21}/n_{2+} = 5/104 = 0.048$, $\hat{p} = (18 + 5)/(189 + 104) = 0.078$, and so

$$z = \frac{0.095 - 0.048}{\sqrt{0.78(1 - 0.078)\left[\frac{1}{189} + \frac{1}{104}\right]}} = 1.435$$

which seems like mild evidence against $H_0 : p_{11} = p_{21}$ (one-sided p-value is 0.076; two-sided p-value is 0.151).

Aside

- Any test based on z =(statistic)/(SE of statistic) is called a *Wald test*.
- You can always invert a Wald Test to produce a confidence interval for the estimand of the numerator, e.g.:

  $$\hat{p}_{11}-\hat{p}_{21}-z_{1-\alpha/2}\sqrt{\mathrm{Var}\,(\hat{p}_{11}-\hat{p}_{21})} \le p_{11}-p_{21} \le \hat{p}_{11}-\hat{p}_{21}+z_{1-\alpha/2}\sqrt{\mathrm{Var}\,(\hat{p}_{11}-\hat{p}_{21})}$$

  Wald intervals are usually the first intervals we teach or use when constructing interval estimates.
- The above Wald interval is reasonably well-behaved in the sense that its nominal coverage is really around $100(1-\alpha)\%$ for most reasonable sample sizes and values of p_{11}, p_{21}.

- *However*, the corresponding Wald interval for a single proportion is quite badly behaved, as I was reminded at a project advisory meeting recently: the nominal coverage of

  $$\hat{p} - z_{1-\alpha/2}\sqrt{\mathrm{Var}\,(\hat{p})} \le p \le \hat{p} + z_{1-\alpha/2}\sqrt{\mathrm{Var}\,(\hat{p})}$$

  may be quite far from $100(1-\alpha)\%$, especially when p is near 0 or 1, since $\sqrt{\mathrm{Var}\,(\hat{p})} = \sqrt{\hat{p}(1-\hat{p})/n}$ may become too small.

  Thus many other intervals have been considered (see e.g. Brown et al., 2001, *Stat. Sci.*, 101–133) and some are even suggested in elementary books (E.g. Moore & McCabe).
- Two of the more successful alternatives are the *Wilson interval*

$$\frac{n\hat{p}+z^2/2}{n+z^2} - \frac{z\sqrt{n}}{n+z^2}\sqrt{\hat{p}(1-\hat{p})+z^2/4n} \le p \le \frac{n\hat{p}+z^2/2}{n+z^2} + \frac{z\sqrt{n}}{n+z^2}\sqrt{\hat{p}(1-\hat{p})+z^2/4n}$$

  and the *Jeffreys interval*

  $$\beta(\alpha/2, n\hat{p}+1/2, n(1-\hat{p})+1/2) \le p \le \beta(1-\alpha/2, n\hat{p}+1/2, n(1-\hat{p})+1/2)$$

  which both continue to behave well when $\hat{p}$ is near 0 or 1.

The Chi-Squared Test

The two-sample test (Wald test) works well for two binomials but doesn't generalize to more than two binomials (rows) or more than two outcomes (columns); nor does it generalize to other sampling schemes (e.g. consider the table as an observational study with only n_{++} fixed).

We can get an idea of a better test by squaring the Wald test:

$$z^2 = \frac{(\hat{p}_{11} - \hat{p}_{22})^2}{\overline{p}(1-\overline{p})\left[\frac{1}{n_{1+}} + \frac{1}{n_{2+}}\right]} = \frac{(n_{11}/n_{1+} - n_{21}/n_{2+})^2}{\frac{(n_{11}+n_{21})(n_{12}+n_{22})}{(n_{1+}+n_{2+})^2} \cdot \frac{n_{1+}+n_{2+}}{n_{1+}n_{2+}}} = \cdots =$$

$$\frac{[n_{11}n_{22} - n_{21}n_{12}]^2 n_{++}}{n_{1+}n_{2+}n_{+1}n_{+2}} = \cdots = \sum_i \sum_j \frac{(n_{ij} - n_{i+}n_{+j}/n_{++})^2}{n_{i+}n_{+j}/n_{++}} = X^2$$

Since z is approx $N(0,1)$, then X^2 is approximately χ^2 on 1 d.f.

For the Aspirin-Heart Attack data, $X^2 = 2.061$, with a p-value of 0.151, consistent with the two-sided z-test above.

Another motivation for the χ^2 test

Now we consider

	Myocardial Infarction		
	Fatal Attack	Nonfatal Attack	Total
Placebo	$n_{11} = 18$	$n_{12} = 171$	$n_{1+} = 189$
Aspirin	$n_{21} = 5$	$n_{22} = 99$	$n_{2+} = 104$
Total	$n_{+1} = 23$	$n_{+2} = 270$	$n_{++} = 293$

as having only the total n_{++} fixed (as in an observational study).

Now we assume $p_{ij} = P(\text{treatment } i \text{ and outcome } j)$.

A test of the effectiveness of the treatment is now a test of

$$H_0 : (\text{treatment}) \perp\!\!\!\perp (\text{outcome})$$

i.e. of whether $p_{ij} = p_{i+}p_{+j}$, or equivalently whether $p_{j|i} = p_{+j}$.

To test $H_0: \ p_{ij} = p_{i+}p_{+j}$, we can compare the table of observed counts n_{ij} with estimates of the expected counts m_{ij} under H_0:

	$j = 1$	$j = 2$
$i = 1$	n_{11}	n_{12}
$i = 2$	n_{21}	n_{22}

vs.

	$j = 1$	$j = 2$
$i = 1$	$\hat{m}_{11}$	$\hat{m}_{12}$
$i = 2$	$\hat{m}_{21}$	$\hat{m}_{22}$

Since $p_{ij} = P(\text{treatment } i \text{ and outcome } j)$, we know that $m_{ij} = n_{++}p_{ij} = n_{++}p_{i+}p_{+j}$ under H_0. Therefore,

$$\hat{m}_{ij} = n_{++}\hat{p}_{i+}\hat{p}_{+j} = n_{++}(n_{i+}/n_{++})(n_{+j}/n_{++}) = n_{i+}n_{+j}/n_{++}$$

Thus we want to compare n_{ij} vs. $n_{i+}n_{+j}/n_{++}$ in each cell of the table. This clearly can be done with the χ^2 statistic derived earlier:

$$X^2 = \sum_i \sum_j \frac{(n_{ij} - n_{i+}n_{+j}/n_{++})^2}{n_{i+}n_{+j}/n_{++}}$$

Aside

Suppose the χ^2 test rejects. What went wrong? A good way to check is by looking at the *Pearson residuals*, which are the square roots of the summands of X^2:

$$\tilde{r}_{ij} = \frac{n_{ij} - \hat{m}_{ij}}{\sqrt{\hat{m}_{ij}}}$$

For the Aspirin-Heart Attack study above, we have

Observed:

n_{ij}	$j = 1$	$j = 2$
$i = 1$	18	171
$i = 2$	5	99

Expected:

$\hat{m}_{ij}$	$j = 1$	$j = 2$
$i = 1$	14.84	174.16
$i = 2$	8.16	95.84

Residuals:

$\tilde{r}_{ij}$	$j = 1$	$j = 2$
$i = 1$	0.82	−0.24
$i = 2$	−1.11	0.32

The Odds Ratio

Another measure of “unrelatedness” (independence) in a 2×2 table is the *odds ratio*.

- For independent binomial rows $p_{ij} = P[j|i]$ so the odds ratio comparing p_{11} to p_{21} is
$$\frac{P[Fatal|Placebo]}{1-P[Fatal|Placebo]} \cdot \frac{1-P[Fatal|Aspirin]}{P[Fatal|Aspirin]} = \frac{p_{11}}{p_{12}} \cdot \frac{p_{22}}{p_{21}} = \frac{p_{11}p_{22}}{p_{12}p_{21}}$$
- For the model in which the total n_{++} is fixed, $p_{ij} = P[i \;\&\; j]$, so the odds ratio to check for independence is
$$\frac{P[Fatal|Placebo]}{1-P[Fatal|Placebo]} \cdot \frac{1-P[Fatal|Aspirin]}{P[Fatal|Aspirin]} = \frac{p_{11}/p_{1+}}{p_{12}/p_{1+}} \cdot \frac{p_{22}/P_{2+}}{p_{21}/P_{2+}} = \frac{p_{11}p_{22}}{p_{12}p_{21}}$$
- Either way, when $p_{11}p_{22}/p_{12}p_{21} = 1$, $P[Fatal|Placebo] = P[Fatal|Aspirin]$, i.e. “Outcome” $\perp\!\!\!\perp$ “Treatment”.
- In both cases,
$$\widehat{O.R.} = \frac{\hat{p}_{11}\hat{p}_{22}}{\hat{p}_{12}\hat{p}_{21}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

For example for the Aspirin-Heart Attack study, $\widehat{O.R.} = 2.084$.

Inference on the Odds Ratio

We will see later in the course that the estimated log-odds-ratio

$$\log \widehat{O.R.} = \log n_{11} + \log n_{22} - \log n_{12} - \log n_{21}$$

has asymptotic standard error

$$SE = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{22}} + \frac{1}{n_{12}} + \frac{1}{n_{21}}}$$

as n_{++} grows (and the true probabilities remain fixed). This immediately gives rise to

- *A Wald test for independence.* In the Aspirin-Heart Attack study, $SE = \sqrt{1/18 + 1/99 + 1/171 + 1/5} = 0.521$, so the test statistic is $z = \log(2.084)/0.521 = 1.409$, so we still do not reject independence.
- *A confidence interval for* $\log$ *OR or for OR itself.* A 95% log O.R. interval would be
$$-0.282 = \log(2.084) - 1.95 \times 0.541 \le \log O.R. \le \log(2.084) + 1.95 \times 0.541 = 1.750$$
Exponentiating both endpoints we get a 95% CI for the O.R. itself:
$$0.755 = \exp(-0.282) \le O.R. \le \exp(1.750) = 5.757$$

$I \times J$ Tables

We return now to the full Aspirin-Heart Attack table.

	Myocardial Infarction			
	Fatal	Nonfatal	No	
	Attack	Attack	Attack	Total
Placebo	$n_{11} = 18$	$n_{12} = 171$	$n_{13} = 10845$	$n_{1+} = 11034$
Aspirin	$n_{21} = 5$	$n_{22} = 99$	$n_{23} = 10993$	$n_{2+} = 11097$
Total	$n_{+1} = 23$	$n_{+2} = 270$	$n_{+3} = 21838$	$n_{++} = 22131$

- Our "independent Binomial rows" model (product-binomial) can be generalized to a *product-multinomial model*: we build independent multinomial distributions for each row of this 2×3 table.

  Under this model, the row-totals n_{i+} are fixed and again $p_{ij} = P[j|i]$. Moreover, $m_{ij} = n_{i+}p_{ij}$.

- Alternatively we can fix only n_{++}, and assume a *single multinomial model* for all six cells in the table. Then $p_{ij} = P[i \,\&\, j]$, and $m_{ij} = n_{++}p_{ij}$.

Chi-squared Test

For a null model $H_0 : \; m_{ij} = m^0_{ij}$, we can form the χ^2 test just as before:

$$X^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - \hat{m}^0_{ij})^2}{\hat{m}^0_{ij}}$$

For large samples (we will formalize this later), X^2 is approximately χ^2 with d.f. depending on I, J and the nature of the null model m^0_{ij}.

For example, when $H_0 : \;$ (rows)$\perp\!\!\!\perp$(columns), we have as before

$$\text{\textit{Product Multinomial}} \quad \Rightarrow \quad p_{1j} = \ldots = p_{Ij} \; \forall j; \;\; \Rightarrow \; \hat{m}^0_{ij} = n_{i+}\hat{p}_{ij} = \frac{n_{i+}n_{+j}}{n_{++}};$$

$$\text{\textit{Single Multinomial}} \quad \Rightarrow \quad p_{ij} = p_{i+}p_{+j} \; \forall i, j; \;\; \Rightarrow \; \hat{m}^0_{ij} = n_{++}\hat{p}_{i+}\hat{p}_{+j} = \frac{n_{i+}n_{+j}}{n_{++}}.$$

and in either case X^2 is asymptotically χ^2 with d.f. $= (I-1)(J-1)$.

Using the full Aspirin-Heart Attack data, we get the table of expected values

$\hat{m}_{ij}$	$j = 1$	$j = 2$	$j = 3$
$i = 1$	11.47	134.62	10887.92
$i = 2$	11.53	135.38	10950.08

and Pearson residuals

$\tilde{r}_{ij}$	$j = 1$	$j = 2$	$j = 3$
$i = 1$	1.93	3.14	−0.41
$i = 2$	−1.92	−3.13	0.41

Then $X^2 = \sum_i \sum_j \tilde{r}_{ij}^2 = 27.27$; with $d.f. = (I-1)(J-1) = 2$, for a p-value of 1.14×10^{-6}.

We can see from the Pearson's residuals that subjects taking Aspirin regularly ($i = 2$) have too few heart attacks ($j = 1, 2$) for independence (H_0) to hold, and subjects on Placebo $i = 1$ have too many. Combining this with our previous analyses:

- Regular aspirin dosage seems to be associated with lower incidence of M.I.'s
- Given that ther is an M.I., regular aspirin dose doesn't appear to affect the severity of the attack.

Odds Ratios

In an $I \times J$ table, there are $\binom{I}{2} \times \binom{J}{2}$ odds ratios

$$OR(i, j, i', j') = \frac{p_{ij}p_{i'j'}}{p_{ij'}p_{i'j}}$$

It is a matter of algebra and patience to show that

- *Under the single multinomial or product multinomial model,* H_0 : *(rows)⫫(columns)* $\Leftrightarrow OR(i, j, i', j') = 1 \;\forall\; i, j, i', j'$.
- $OR(i, j, i', j') = 1 \;\forall\; i, j, i', j'$, *if and only if* $OR(1, 1, i, j) = 1 \;\forall\; i, j$.

Thus (a) the OR's can be used to explore deviations from independence (much like Pearson residuals); and (b) there is a lot of redundancy in OR's for $I \times J$ tables!

Note: We can estimate $\widehat{OR}(i, j, i', j')$ and formulate Wald tests and intervals for it, using the same formulae that work for 2×2 tables.

In the Aspirin-Heart Attack data,

n_{ij}	$j=1$	$j=2$	$j=3$
$i=1$	18	171	10845
$i=2$	5	99	10993

there are several interesting odds ratios to consider:

i, j, i', j'	OR	$z_{H_0:\ OR=1}$	p-value	CI for OR
1,1,2,2	2.08	1.41	0.16	0.75 to 5.79
1,1,2,3	3.65	2.56	0.01	1.35 to 9.83
1,2,2,3	1.75	4.41	0.00	1.37 to 2.25

- From $OR(1,1,2,2)$ we see again that Aspirin and Fatality are not related;
- From $OR(1,1,2,3)$ and $OR(1,2,2,3)$ we see that subjects taking aspirin have fewer heart attacks of each kind separately, than subjects who do not, relative to the independence model.

Another interesting table would pool across the two M.I. conditions:

n_{ij}	$j=1$	$j=2$
$i=1$	189	10845
$i=2$	104	10993

$\tilde{r}_{ij}$	$j=1$	$j=2$
$i=1$	0.82	−0.24
$i=2$	−1.11	0.32

For this table, $X^2 = 24.89$ on 1 d.f., with a p-value of 6.06×10^{-7}; $\widehat{OR} = 1.84$ with $z_{H_0:\ OR=1} = 4.97$. The Pearson residuals above again show that Aspirin and M.I. avoidance are positively associated.

Questions for the future:

- When is pooling (adjacent) categories justified?
- When does it lead to the “same” answer as not pooling?

For now though, we take a closer look at the *sampling models* underlying most contingency table analysis.

Sampling Models

- No fixed totals: *Poisson*
  Cross-sectional study: Cull auto-crash records and record, for each person, whether seat belt was worn (i), and whether crash was fatal (j). If the number of records is limited only by time (or money) for the study: Poisson.
- Total $n = n_{++}$ fixed: *Multinomial*
  In the auto crash study, if the total number of records to be culled is fixed (say, $n = 100$): Multinomial.
- Fixed rows (or columns): *Product Multinomial*
  – *Retrospective, or case-control study:* Sample a fixed number of cancer and non-cancer patients, and look backwards to see who smoked.
  – *Prospective studies:* Sample a fixed number of treated and non-treated individuals and follow them until an outcome occurs
    ∗ *Clinical trials*: experimenter assigns tx's
    ∗ *Cohort studies*: subjects self-select tx's
- (Fixed rows and columns: *Hypergeometric*)

Some Sampling Theory

We consider a table of counts $n_1, \ldots, n_r$ with associated probabilities $p_1, \ldots, p_r$. We also let $n = \sum_{i=1}^{r} n_i$.

We will first consider the *product multinomial model*:

- Re-index the n's and p's to be n_{ij}, p_{ij} with $i = 1, \ldots, I$ and $j = 1, \ldots, J$.
- The model is then

$$L(p) = \prod_{i=1}^{I} \left[\frac{n_{i+}!}{\prod_{j=1}^{J} n_{ij}!} \prod_{j=1}^{J} p_{ij}^{n_{ij}} \right]$$

We wish to find the MLE $\hat{p} = (\hat{p}_1, \ldots, \hat{p}_r)$. It is easiest to consider the log-likelihood

$$\ell(p) = \log L(p) = \sum_{i=1}^{I} \left[\log(n_{i+}!) - \sum_{j=1}^{J} \log(n_{ij}!) + \sum_{j=1}^{J} n_{ij} \log p_{ij} \right]$$

The only term involving p is $\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}\log p_{ij}$. The maximum is achieved when we maximize each term

$$f_i(p) = \sum_{j=1}^{J} n_{ij}\log p_{ij}$$

To maximize this we rely on the follwing lemma:

Lemma *Let $f(p_1,\ldots,p_r) = \sum_{i=1}^{r} n_i \log p_i$, s.t. $p_i \geq 0$, $n_i > 0$, and $\sum_i p_i = 1$. Then $f()$ is maximized by $\hat{p}_i = n_i/n$, where $n = \sum_i n_i$.*

Proof. Lagrange Multiplier problem, with objective function

$$g(p,\lambda) = f(p_1,\ldots,p_r) - \lambda\left(\sum_i p_i - 1\right) \qquad \square$$

Therefore, the MLE's for the *saturated/unrestricted product multinomial* are

$$\hat{p}_{ij} = n_{ij}/n_{i+}$$

and we get corresponding ML expected counts $\hat{m}_{ij} = n_{i+}\hat{p}_{ij} = n_{ij}$ (as we might expect!).

If we impose a restriction, say H_0, on the p_{ij}'s (or equivalently the m_{ij}'s) we must incorporate that restriction in the maximization, producing new MLE's $\hat{p}^0_{ij}$ and $\hat{m}^0_{ij}$.

For example if the restriction is $H_0 : p_{1j} = \cdots = p_{Ij},\ \forall j$, we may set $\pi_j = p_{1j} = \cdots = p_{Ij}$, and the new loglikelihood is

$$\ell(p) = \log L(p) = \sum_{i=1}^{I}\left[\log(n_{i+}!) - \sum_{j=1}^{J}\log(n_{ij}!) + \sum_{j=1}^{J} n_{ij}\log\pi_j\right]$$

and the term involving π_i's is now

$$\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}\log\pi_j = \sum_{j=1}^{J} n_{+j}\log\pi_j$$

If we apply the MLE lemma again we get

$$\hat{p}^0_{ij} = \hat{\pi}_j = n_{+j}/n_{++}$$

and $\hat{m}^0_{ij} = n_{i+}\hat{p}^0_{ij} = n_{i+}n_{+j}/n_{++}$, just as we expect.

Likelihood Ratio Test

What does the Likelihood Ratio Test look like? We want to consider $-2[\log L(\hat{p}^0) - \log L(\hat{p})]$. Our two maximized log-likelihoods are

$$\begin{aligned}
\log L(p^0) &= \sum_{i=1}^{I}\left[\log(n_{i+}!) - \sum_{j=1}^{J}\log(n_{ij}!) + \sum_{j=1}^{J} n_{ij}\log(n_{+j}/n_{++})\right] \\
\log L(p) &= \sum_{i=1}^{I}\left[\log(n_{i+}!) - \sum_{j=1}^{J}\log(n_{ij}!) + \sum_{j=1}^{J} n_{ij}\log(n_{ij}/n_{i+})\right]
\end{aligned}$$

Taking the difference, multiplying by -2 and simplifying, we get

$$G^2 = 2\sum_{i=1}^{I}\sum_{j=1}^{J}\hat{m}_{ij}\log\left(\frac{\hat{m}_{ij}}{\hat{m}^0_{ij}}\right)$$

where $\hat{m}_{ij} = n_{ij}$ and $\hat{m}^0_{ij} = n_{i+}n_{+j}/n_{++}$.

It is well known that the LR statistic (G^2 in our case) is asymptotically χ^2 under H_0, with df equal to the number of linear restrictions getting from H_A : “m_{ij} unrestricted”, to H_0.

In our case, $G^2 \sim \chi^2$ with $d.f. = (I-1)(J-1)$, just as for X^2, and G^2 can be used in the same way:

For example, for the full 2×3 Aspirin-Heart Attack table, we obtained $X^2 = 27.27$; with $d.f. = (I-1)(J-1) = 2$, for a p-value of $1.14{\times}10^{-6}$

Using G^2 instead we obtain $G^2 = 28.06$; with $d.f. = (I-1)(J-1) = 2$, for a p-value of 8.08×10^{-7}. This rejection of independence is just as strong as with X^2.