Combining Results From Multiple 2×2 Tables Motivating Logistic Regression

Statistics 149

Spring 2006

Excess in 2 $\times$ 2 Tables

As mentioned last time, for tables with fixed margins, under the $H_0 : \omega_1 = \omega_2$,

$$n_{11} \sim Hyper(R_1, C_1, T)$$

The mean and variance of this hypergeometric are

$$E[n_{11}] = \frac{R_1 C_1}{T} \qquad \mathrm{Var}(n_{11}) = \frac{R_1 R_2 C_1 C_2}{T^2(T-1)}$$

So another measure of how much a particular dataset deviates from the null hypothesis is

$$Excess = Observed - Expected$$

It doesn't matter which cell you calculate this for as

$$|Excess_{ij}| = |O_{ij} - E_{ij}|$$

and

$$\begin{aligned} \mathrm{Var}(Excess_{ij}) &= \mathrm{Var}(n_{ij}) \\ &= \frac{R_1 R_2 C_1 C_2}{T^2(T-1)} \end{aligned}$$

is the same for all four cells in a 2 $\times$ 2 table.

So an alternative test is to use the test statistic

$$z = \frac{Excess}{\sqrt{\mathrm{Var}(Excess)}}$$

When the sample sizes are large, the sampling distribution of z is approximately $N(0, 1)$.

This gives alternatives to Fisher's Exact and Pearson's Chi-square tests. However it is usually used for different purposes, the comparison of multiple 2×2 tables.

Comparing N Odds Ratios

Example: Ille-et-Vilaine Study of Oesophageal Cancer

A retrospective study to examine the relationship with alcohol consumption and oesophageal cancer.

Alcohol	Cancer	No Cancer	Total
High	96	109	205
Low	104	666	770
Total	200	775	975

There is an additional complication in that subjects range from 25 years in age to older than 75. Age is strongly associated with cancer with the older somebody is, the more likely they are to have the disease.

	Cancer		No Cancer	
	High Alcohol	Low Alcohol	High Alcohol	Low Alcohol
Age	n_{11}	n_{12}	n_{21}	n_{22}
25-34	1	0	9	106
35-44	4	5	26	164
45-54	25	21	29	138
55-64	42	34	27	139
65-74	19	36	18	88
75+	5	8	0	31
Total	96	104	109	666

So instead of asking whether the odds of cancer is the same for both alcohol consumption levels, it makes more sense to ask whether the odds of cancer is the same for both alcohol consumption levels within each age group.

So there may be different rates of alcohol consumption (the response variable in this retrospective study) across the different age groups, but is the pattern of alcohol consumption the same between the cancer and non cancer groups within each age group.

	Cancer		No Cancer		
	High Alcohol	Low Alcohol	High Alcohol	Low Alcohol	
Age	n_{11}	n_{12}	n_{21}	n_{22}	$\hat{\phi}_k$
25-34	1	0	9	106	∞
35-44	4	5	26	164	4.98
45-54	25	21	29	138	5.61
55-64	42	34	27	139	6.30
65-74	19	36	18	88	2.56
75+	5	8	0	31	∞
Total	96	104	109	666	

We are interested in

$$H_0 : \phi_1 = \phi_2 = \ldots = \phi_N = 1$$

Are the odds rates in each strata $= 1$?

Since all of the estimated odds seem to be far from 1, it appears that the data in this case don't support the null hypothesis. But we want a better test to examine this observation.

Lets think what happens in table k. We can use the $Excess$ in this table to examine whether $\phi_k = 1$ or not.

In addition, under reasonable sampling schemes, the $Excess$ measures in each subtable is independent of those in the other tables.

One way to combine the information from each subtable to get an overall summary on the null hypothesis is to add the $Excess$ from each 2×2 table.

Mantel-Haenszel Test

$$Y = \sum_{k=1}^{N} Excess_k$$

Under the null hypothesis

$$E[Y] = 0$$

and

$$\begin{aligned} \mathrm{Var}(Y) &= \sum_{k=1}^{N} \mathrm{Var}(Excess_k) \\ &= \sum_{k=1}^{N} \frac{R_{1k} R_{2k} C_{1k} C_{2k}}{T_k^2 (T_k - 1)} \end{aligned}$$

where R_{ik}, C_{jk}, and T_k are the row, column, and grand totals from table k.

The usual test statistic to examine this is

$$z = \frac{Y}{\sqrt{\mathrm{Var}(Y)}}$$
$$= \frac{\sum Excess_k}{\sqrt{\sum \frac{R_{1k}R_{2k}C_{1k}C_{2k}}{T_k^2(T_k-1)}}}$$

which is compared to a $N(0,1)$ distribution. This is known as the Mantel-Haenszel Test (sometimes as Cochran-Mantel-Haenszel Test).

Note that this is an asymptotic result. A common rule of thumb is based on summing the expected counts across the tables, i.e. look at

$$\begin{array}{cc} E_{11+} & E_{12+} \\ E_{21+} & E_{22+} \end{array}$$

If each of the $E_{ij+} \geq 5$, the normal approximation shouldn't be too bad.

	Cancer		No Cancer				
	High	Low	High	Low			
Age	n_{11}	n_{12}	n_{21}	n_{22}	Expected	Excess	Variance
25-34	1	0	9	106	0.086	0.914	0.079
35-44	4	5	26	164	1.357	2.643	1.106
45-54	25	21	29	138	11.662	13.338	6.858
55-64	42	34	27	139	21.669	20.331	10.671
65-74	19	36	18	88	12.640	6.360	6.449
75+	5	8	0	31	1.477	3.523	0.944
Total					48.891	47.109	26.106

$$z = \frac{47.109}{\sqrt{26.106}} = 9.22$$

$$p\text{-value} = 2P[Z \geq 9.22] = 2.9 \times 10^{-20}$$

In this example, just over 47 more cancer patients were high alcohol consumers than would be expected if the odds of cancer were the same for high and low alcohol subjects. The p-value suggests that this result is highly statistically significant (and practically as well).

So it appears that alcohol consumption is associated with oesophageal cancer.

Often a Chi-squared version of the Mantel-Haenszel test will be used instead of what has been discussed. The test statistic is

$$X^2 = \frac{Y^2}{\mathrm{Var}(Y)} = \frac{(\sum Excess_k)^2}{\sum \frac{R_{1k}R_{2k}C_{1k}C_{2k}}{T_k^2(T_k-1)}} = z^2$$

and is compared to a χ_1^2 distribution.

This is the version that **R** presents in the function `mantelhaen.test`

```
> cancer <- array(
+   c( 1,  9,  0, 106,
+      4, 26,  5, 164,
+     25, 29, 21, 138,
+     42, 27, 34, 139,
+     19, 18, 36,  88,
+      5,  0,  8,  31), c(2,2,6),
+   dimnames = list(Cancer=c("Yes","No"), Alcohol=c("High","Low"),
+     Age=c("25-34","35-44","45-54","55-64","65-74","75+")))
>
> cancer
, , Age = 25-34

      Alcohol
Cancer High Low
   Yes    1   0
   No     9 106
```

```
, , Age = 35-44

      Alcohol
Cancer High Low
   Yes    4   5
   No    26 164

> mantelhaen.test(cancer, correct=F)

   Mantel-Haenszel chi-squared test without continuity correction

data:  cancer
Mantel-Haenszel X-squared = 85.0095, df = 1, p-value < 2.2e-16
alternative hypothesis: true common odds ratio is not equal to 1
95 percent confidence interval:
 3.562131 7.467743
sample estimates:
common odds ratio
         5.157623
```

Note the default of `mantelhaen.test(array, alternative="two.sided")` is a two-sided test, which compares the hypotheses

$$H_0 : \phi_1 = \ldots = \phi_N = 1 \qquad \text{vs} \qquad H_A : \phi_1 = \ldots = \phi_N \neq 1$$

It is also possible to do one sided tests which deal with hypotheses

- `mantelhaen.test(array, alternative="less")`

$$H_0 : \phi_1 = \ldots = \phi_N = 1 \qquad \text{vs} \qquad H_A : \phi_1 = \ldots = \phi_N < 1$$

- `mantelhaen.test(array, alternative="greater")`

$$H_0 : \phi_1 = \ldots = \phi_N = 1 \qquad \text{vs} \qquad H_A : \phi_1 = \ldots = \phi_N > 1$$

This function will also estimate the common odds ratio, assuming that each table is generated by a common odds ratio. The estimate of this common odds ratio is

$$\hat{\phi} = \frac{\sum_{\text{all tables}} n_{11}n_{22}/T}{\sum_{\text{all tables}} n_{12}n_{21}/T}$$

So the estimate of this in the cancer example is

$$\hat{\phi} = \frac{\frac{(1)(106)}{116} + \frac{(4)(164)}{199} + \frac{(25)(138)}{213} + \frac{(42)(139)}{242} + \frac{(19)(88)}{161} + \frac{(5)(31)}{44}}{\frac{(0)(9)}{116} + \frac{(5)(26)}{199} + \frac{(21)(29)}{213} + \frac{(34)(27)}{242} + \frac{(36)(18)}{161} + \frac{(8)(0)}{44}} = 5.16$$

So this implies that the odds of cancer is 5 times more for people with high alcohol consumption than those with low consumption.

An underlying assumption of this test and the estimate of the common odds ratio is that the odds ratio is the same for each table. If this is not true, this test may not act as expected, and it isn't clear what $\hat{\phi}$ is actually estimating. Lets look at an example where the constant odds ratio assumption seems to breakdown.

Example: Graduate Admissions at UC Berkeley

The following is the graduate admissions data at UC Berkeley for their 6 largest graduate programs in 1973 (available in **R** in the array `UCBAdmissions`). One of the initial concerns was whether there was any gender discrimination in this data as

Gender	Admitted	Rejected	% Admitted
Male	1198	1493	44.5
Female	557	1278	30.6

One thing that needs to be noted that there are different admission rates for the different programs

Major	Admitted	Rejected	Applied	% Admit	% Reject
A	600	333	933	64.31	35.69
B	370	215	585	63.25	36.75
C	322	596	918	35.08	64.92
D	269	523	792	33.96	66.04
E	148	436	584	25.34	74.66
F	46	668	714	6.44	93.56
Total	1755	2771	4526	38.78	61.22

So we should take a look at happens within each major separately.

	Men		Women				
	Admit	Reject	Admit	Reject			
Major	n_{11}	n_{12}	n_{21}	n_{22}	$\hat{\phi}_k$	Men %	Women %
A	512	313	89	19	0.35	62	82
B	353	207	17	8	0.80	63	68
C	120	205	202	291	1.13	37	34
D	138	279	121	244	0.92	33	35
E	53	138	94	299	1.22	28	24
F	22	351	24	317	0.83	6	7

Looking at things this way, it appears that there isn't really evidence of discrimination against women. Women seem to do about the same as or better (major A) than men. When looking at the aggregate 2×2 table, the fact that women tend to apply to the more difficult programs for admittance affects the observed relationship.

This data set is an example Simpson's Paradox

> An association or comparison that holds for all of several groups can reverse direction when the data are combined to form a single group.

If we look at just the 2×2 table, $\hat{\phi} = 1.80$. But this is driven by about half the men applying to majors A and B, but only about 5% of the women.

Note that the odds ratio in the aggregated table is much larger than any odds ratio in the subtables.

In this example, major is an example of a lurking variable, a variable that has (potentially) an important effect, but it ignored in an analysis.

Lets look at the Mantel-Haenszel test on this dataset

```
> mantelhaen.test(UCBAdmissions, correct=F)

   Mantel-Haenszel chi-squared test without continuity correction

data:  UCBAdmissions
Mantel-Haenszel X-squared = 1.5246, df = 1, p-value = 0.2169
alternative hypothesis: true common odds ratio is not equal to 1
95 percent confidence interval:
 0.7719074 1.0603298
sample estimates:
common odds ratio
        0.9046968
```

So this test suggests there is nothing going on with the relationship between admissions and gender, missing what is going on in Major A.

It is possible to show that in this example the odds ratio isn't constant by Woolf's test for interaction (p-value $= 0.003$).

While there is an interaction in admissions data, in many cases a constant odds ratio assumption is reasonable. Another way of thinking of this is while the odds and probabilities for success can very greatly across different tables, the odds ratio for each table are approximately the same. For example, the cancer example shows this

	Cancer		No Cancer		
	High Alcohol	Low Alcohol	High Alcohol	Low Alcohol	
Age	n_{11}	n_{12}	n_{21}	n_{22}	$\hat{\phi}_k$
25-34	1	0	9	106	∞
35-44	4	5	26	164	4.98
45-54	25	21	29	138	5.61
55-64	42	34	27	139	6.30
65-74	19	36	18	88	2.56
75+	5	8	0	31	∞

There isn't much evidence of different odds ratios (p-value $= 0.23$). Note that the tables that appear to deviate have fairly small sample sizes, so the estimates of ϕ_k in those tables is highly variable.

One other issue with the Mantel-Haenszel Test is it treats the confounding variable as nominal. While this is reasonable for the admissions example, it potentially misses the ordering of age in the cancer example.

This is more important with modeling the odds ratio. For example we could think of fitting something like

$$\phi(age) = \alpha + \beta age$$

or

$$\log(\phi(age)) = \alpha + \beta age$$

Logistic Regression Motivation

In the previous examples, we have been looking modeling success probabilities or odds based on categorical variables. For example

- $P[\text{Stroke}|\text{Treatment}]$
- $P[\text{Lives one year}|\text{Birth Weight}]$
- $P[\text{Admission}|\text{Gender, Major}]$

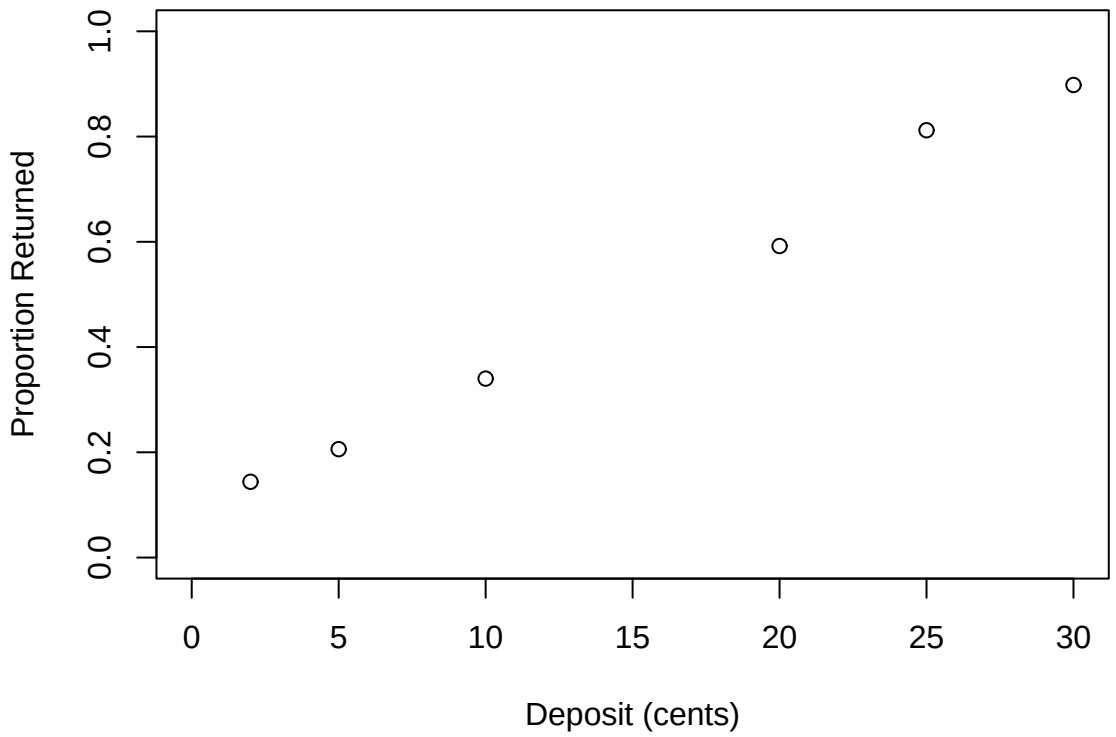
In all of these examples, the predictor variables are discrete, or at least treated that way.

There are examples where the predictor variables are continuous (or having enough levels where treating them as continuous is reasonable)

Example: Bottle Return

A carefully controlled experiment was conducted to study the effect of the size of the deposit level on the likelihood that a returnable one-litre soft-drink bottle will be returned. The following data show the number of bottles returned (y_i) out of 500 sold (n_i) at each of 6 deposit levels (x_i, in cents)

Observation i:	1	2	3	4	5	6
Deposit level x_i:	2	5	10	20	25	30
Number sold n_i:	500	500	500	500	500	500
Number returned y_i:	72	103	170	296	406	449
Prop. returned:	0.144	0.206	0.340	0.592	0.812	0.898



So it is fairly clear that increasing the deposit that a person needs to pay, the more likely a bottle is returned.

So we want to derive a model for the probability that a bottle that a bottle is returned, π, given the deposit x.

Let Z_i be the response for the ith bottle where

$$Z_i = \begin{cases} 1 & \text{Returned} \\ 0 & \text{Not returned} \end{cases}$$

Then

$$Z_i \sim Bin(1, \pi(x_i))$$

where x_i is the deposit for bottle i.

An equivalent way of thinking of this, is to model

$$\mu(Z_i|x_i) = \pi(x_i)$$

This is the analogue to linear regression, where we are trying to describe

$$\mu(Y_i|x_i) = f(x_i) \quad (= \beta_0 + \beta_1 x_i \text{ often})$$

One possible model would be

$$\pi(x_i) = \beta_0 + \beta_1 x_i$$

Lets fit the observed data $(x_i, \hat{p}(x_i))$ by least squares. Lets ignore that this is suboptimal since the variances $\hat{p}(x_i)$ can be constant.

```
> return.lm <- lm(ret.prop[,1] ~ deposit)

> summary(return.lm)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.071542   0.022413   3.192   0.0332 *
deposit     0.027856   0.001211  22.996 2.12e-05 ***
```

Now lets see what the model predicts the return probabilities to be under this model

```
> pred.levels <- data.frame(deposit=seq(30,50,5))
> predict(return.lm, pred.levels)
        1         2         3         4         5
0.9072207 1.0465005 1.1857803 1.3250601 1.4643399
```

So for deposits over 35 cents (actually 33.33 cents) will have estimated predict return probabilities of greater than 1.

While it doesn't make much sense here, what would we predict if $x = -5$ (You have to pay to recycle the bottle). In this case, the estimated probability is -0.0677.

So this model has problems in that it can give probabilities outside [0,1]. So we need a different model for modeling binary responses.

The approach we will take next class is to model π as

$$g(\pi) = \beta_0 + \beta_1 x$$

for a nice function g. We want to choose g such that $g(\pi)$ can range from $(-\infty, \infty)$, the possible range of $\beta_0 + \beta_1 x$.