Sections 6.4, 6.5, 6.6

Timothy Hanson

Department of Statistics, University of South Carolina

Stat 770: Categorical Data Analysis

6.4 $2 \times 2 \times K$ tables

Clinical trial w/ 8 centers, two creams compared to cure infection (p. 226).

Center $Z = k$	Treatment X	Response Y		$\widehat{\theta}_{XY(k)}$
		Success	Failure	
1	Drug	11	25	1.2
	Control	10	27	
2	Drug	16	4	1.8
	Control	22	10	
3	Drug	14	5	4.8
	Control	7	12	
4	Drug	2	14	2.3
	Control	1	16	
5	Drug	6	11	∞
	Control	0	12	
6	Drug	1	10	∞
	Control	0	10	
7	Drug	1	4	2.0
	Control	1	8	
8	Drug	4	2	0.3
	Control	6	1	

6.4.1 Same setup as Section 2.3

Have:

- Y binary outcome (e.g. success/failure of treatment).
- X binary predictor (e.g. treatment).
- Stratum Z (e.g. treatment center).

Want to test $X \perp Y|Z$ versus an alternative. Let $\pi_{ik} = P(Y = 1|X = i, Z = k)$ and

$$\theta_{XY(k)} = \frac{P(Y = 1|X = 1, Z = k)/P(Y = 2|X = 1, Z = k)}{P(Y = 1|X = 2, Z = k)/P(Y = 2|X = 2, Z = k)}.$$

Recall $X \perp Y|Z$ when $\theta_{XY(k)} = 1$. This happens under the model

$$\text{logit } \pi_{ik} = \alpha + \beta_k^Z.$$

This is an ANOVA-type specification where instead of listing $K - 1$ dummy variables, we concisely include a subscript on Z's effect β_k^Z. So there are K effects for Z, $\beta_1^Z, \beta_2^Z, \ldots, \beta_K^Z$ and they sum to zero.

Two models

An additive alternative model specifies

$$\text{logit } \pi_{ik} = \alpha + \beta I\{i = 1\} + \beta_k^Z.$$

Under this model $\theta_{XY(k)} = e^{\beta}$ for all k. The odds *ratios* are the same across strata, but the strata-specific probabilities of success change with $Z = k$. $X \perp Y|Z$ if we accept $H_0 : \beta = 0$.

The most general alternative is

$$\text{logit } \pi_{ik} = \alpha + \beta I\{i = 1\} + \beta_k^Z + \beta_k^{XZ} I\{i = 1\}.$$

This is a saturated model and allows
$\theta_{XY(1)} \neq \theta_{XY(2)} \neq \cdots \neq \theta_{XY(K)}$. $X \perp Y|Z$ if we accept
$H_0 : \beta = 0, \beta_k^{XZ} = 0$ for $k = 1, \ldots, K$.

Both of these alternatives allow testing $H_0 : X \perp Y|Z$ in PROC LOGISTIC with a Wald test.

6.4.2 Cochran-Mantel-Haenszel statistic

$$\text{CMH} = \frac{\left[\sum_{k=1}^{K}(n_{11k} - \hat{\mu}_{11k})\right]^2}{\sum_{k=1}^{K}\text{var}(n_{11k})},$$

where $\hat{\mu}_{11k} = n_{1+k}n_{+1k}/n_{++k}$ and
$\text{var}(n_{11k}) = n_{1+k}n_{2+k}n_{+1k}n_{+2k}/n_{++k}^2(n_{++k-1})$.

- Motivated by retrospective studies, e.g. case-control, so response (column) totals are assumed fixed. Then row (treatment) totals are sufficient and conditioned on. Leaves only one free parameter in each table, say n_{11k} which is hypergeometric under H_0:
- Null hypothesis is $H_0 : X \perp Y|Z$.
- $\hat{\mu}_{11k} = E(n_{11k})$ and $\text{var}(n_{11k})$ are under H_0.
- When H_0 true, $\text{CMH} \overset{\bullet}{\sim} \chi_1^2$.

A bit more detail why n_{11k} are hypergeometric...

	$Y=1$	$Y=2$	
$X=1$	n_{11k}	n_{12k}	n_{1+k}
$X=2$	n_{21k}	n_{22k}	n_{2+k}
	n_{+1k}	n_{+2k}	n_{++k}

- There are n_{1+k} "red balls" $X=1$ and n_{2+k} "green balls" $X=2$.
- We choose n_{+1k} balls (controls $Y=1$) from the urn. Under independence one cannot tell the difference between a case and a control. The number n_{11k} out of n_{+1k} that are "red," i.e. exposures $X=1$, is hypergeometric (under H_0).
- See page 91, (3.17) in Section 3.5.1.

Back to logistic regression formulation...

The additive alternative looks in a certain direction for deviations from conditional independence $X \perp Y|Z$. It can be more powerful when the additive model truly holds.

The interaction, saturated model can be more powerful when the additive alternative does not hold.

The CMH test is equivalent to a score test for testing $H_0 : \beta = 0$ in the additive model; see your book (p. 227). This test can be carried out in PROC FREQ.

```
data cmh;
input center $ treat response count;
datalines;
a 1 1 11
a 1 2 25
a 2 1 10
a 2 2 27
b 1 1 16
b 1 2 4
...
h 1 1 4
h 1 2 2
h 2 1 6
h 2 2 1
;
proc freq; weight count; tables center*treat*response / cmh;
```

Partial output

Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

Statistic	Alternative Hypothesis	DF	Value	Prob
1	Nonzero Correlation	1	6.3841	0.0115
2	Row Mean Scores Differ	1	6.3841	0.0115
3	General Association	1	6.3841	0.0115

Estimates of the Common Relative Risk (Row1/Row2)

Type of Study	Method	Value	95% Confidence Limits	
Case-Control (Odds Ratio)	Mantel-Haenszel	2.1345	1.1776	3.8692
	Logit **	1.9497	1.0574	3.5949
Cohort (Col1 Risk)	Mantel-Haenszel	1.4245	1.0786	1.8812
	Logit **	1.2194	0.9572	1.5536
Cohort (Col2 Risk)	Mantel-Haenszel	0.8129	0.6914	0.9557
	Logit	0.8730	0.7783	0.9792

** These logit estimators use a correction of 0.5 in every cell of those tables that contain a zero.

We see CMH$=6.384$ with $p = 0.0115$ and so we reject that $X \perp Y|Z$ in favor of a *common odds ratio* estimated as $\hat{\theta}_{XY} = 2.13\ (1.18, 3.87)$.

Testing through logistic regression

Alternatively, we can fit the three logit models:

```
data cmh2;
 input center $ treat y n; treat=abs(treat-2);
 datalines;
 a 1 11 36
 a 2 10 37
 b 1 16 20
 b 2 22 32
...
 h 1 4  6
 h 2 6  7
;
proc logistic data=cmh2; class center; model y/n = center;
proc logistic data=cmh2; class center; model y/n = treat center;
proc logistic data=cmh2; class center; model y/n = treat center treat*center;
```

Label the models (1), (2), and (3) respectively. The fit of (2) corresponds to the alternative in the CMH test:

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
treat	1	6.4174	0.0113
center	7	58.4897	<.0001

Testing through logistic regression

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-1.2554	0.2692	21.7413	<.0001
treat		1	0.7769	0.3067	6.4174	0.0113
center	a	1	-0.0667	0.3133	0.0453	0.8315
center	b	1	1.9888	0.3556	31.2789	<.0001
center	c	1	1.0862	0.3596	9.1236	0.0025
center	d	1	-1.4851	0.5707	6.7711	0.0093
center	e	1	-0.5866	0.4582	1.6390	0.2005
center	f	1	-2.2136	0.9171	5.8260	0.0158
center	g	1	-0.8644	0.7016	1.5178	0.2180

We reject $H_0 : \beta = 0$ ($p = 0.0113$) and thus reject $X \perp Y|Z$. We estimate the common odds ratio to be $e^{-0.777} = 2.18$ $(1.19, 3.97)$ (from excised output).

By adding `/ aggregate scale=none;` to the MODEL statement, we find the Pearson GOF $X^2 = 8.03$ on $df = 16 - (1 + 1 + 7) = 7$ with $p = 0.33$. The additive model does not show gross LOF.

Let's examine the full interaction (saturated) model anyway...

Testing through logistic regression

The `-2 Log L` from (1) is 283.689 (under `Model Fit Statistics`) and from (3) is 267.274. The number of parameters added to (1) to get (3) is 8. The p-value is $P(\chi^2_8 > 16.415) = 0.0368$.

We reject that $H_0 : \beta = 0, \beta_k^{XY} = 0$ in the saturated model (3) and hence also reject $X \perp Y|Z$. Notice the p-value is about 3 times larger though; we lost some power by considering a *very* general alternative.

By accepting this more complex alternative we have lost interpretability as well, the estimated odds ratio $\hat{\theta}_{XY(k)}$ changes with center k. From (3)'s fit `Type 3 Analysis of Effects`:

Effect	DF	Wald Chi-Square	Pr > ChiSq
treat	1	0.0064	0.9362
center	7	24.2036	0.0010
treat*center	7	4.0996	0.7682

The Type III effects table shows we can drop the treat*center from the model and so we go with the analysis and results from the CMH analysis and/or logit analysis on the previous slide.

6.5 Existence of finite $\hat{\beta}$

Estimates $\hat{\beta}$ exist except when data are perfectly separated.

Complete separation happens when a linear combination of predictors perfectly predicts the outcome. See Figure 6.5 (p. 234). Here, there are an infinite number of perfect fitting curves that have $\alpha = \infty$. Essentially, there is a value of x that perfectly separates the 0's and 1's. In two-dimensions there would be a line separating the 0'1 and 1's.

Quasi-complete separation happens when there's a line that separates 0's and 1's but there's some 0's and 1's on the line. We'll look at some pictures.

The end result is that the model will appear to fit but the standard errors will be absurdly large. This is the *opposite* of what's really happening, that the data can be perfectly predicted.

A (Bayesian!) fix is hiding in Section 7.4.7 (p. 275). Add FIRTH to the MODEL statement, and quasi and complete separation issues vanish!

6.6: Power and sample size*

Recall: $\alpha = P(\text{reject } H_0 | H_0 \text{ true})$ and $\beta = P(\text{accept } H_0 | H_1 \text{ true})$.

Power is $1 - \beta = P(\text{reject } H_0 | H_1 \text{ true})$. Often we want to find an overall sample size n such that, for example, $1 - \beta = 0.9$ while capping off $\alpha = 0.05$.

One sample proportion

Say we want to test $H_0 : \pi = \pi_0$ for $Y \sim \text{bin}(n, \pi)$. The score test statistic is $Z_0 = \frac{\hat{\pi} - \pi_0}{\sigma_0}$ where $\hat{\pi} = Y/n$ and $\sigma_0 = \sqrt{\pi_0(1-\pi_0)/n}$. Under $H_0 : \pi = \pi_0$, $Z \overset{\bullet}{\sim} N(0,1)$; this determines $z_{\alpha/2}$. The power $1 - \beta$ is a function of the hypothesized π_0, the true π_1, and the sample size through σ_0 and $\sigma_1 \sqrt{\pi_1(1-\pi_1)/n}$.

Computing the power

$$
\begin{aligned}
1-\beta &= P(\text{reject } H_0 | H_1 \text{ true}) \\
&= P(|Z_0| > z_{\alpha/2} | \pi = \pi_1) \\
&= 1 - P(-z_{\alpha/2} \leq Z_0 \leq z_{\alpha/2} | \pi = \pi_1) \\
&= 1 - P(-z_{\alpha/2}\sigma_0 + \pi_0 \leq \hat{\pi} \leq z_{\alpha/2}\sigma_0 + \pi_0 | \pi = \pi_1) \\
&= 1 - P\left(\frac{-z_{\alpha/2}\sigma_0 + \pi_0 - \pi_1}{\sigma_1} \leq \frac{\hat{\pi} - \pi_1}{\sigma_1} \leq \frac{z_{\alpha/2}\sigma_0 + \pi_0 - \pi_1}{\sigma_1} \right) \\
&= 1 - P\left(\frac{-z_{\alpha/2}\sigma_0 + \pi_0 - \pi_1}{\sigma_1} \leq Z \leq \frac{z_{\alpha/2}\sigma_0 + \pi_0 - \pi_1}{\sigma_1} \right) \\
&= 1 - \left[\Phi\left(\frac{z_{\alpha/2}\sigma_0 + \pi_0 - \pi_1}{\sigma_1} \right) - \Phi\left(\frac{-z_{\alpha/2}\sigma_0 + \pi_0 - \pi_1}{\sigma_1} \right) \right].
\end{aligned}
$$

For a given β, α, π_0, and π_1, we can solve this equation for the sample size n. Check out http://www.cs.uiowa.edu/~rlenth/Power/

6.6.1 Testing $H_0 : \pi_1 = \pi_2$ from two samples

Recall the two-sample proportion problem. Assume the same number of observations n will be collected in each group $X = 1$ and $X = 2$.

$$Y_1 \sim \text{bin}(n_1, \pi_1) \perp Y_2 \sim \text{bin}(n_2, \pi_2).$$

Let $\hat{\pi}_1 = Y_1/n$ and $\hat{\pi}_2 = Y_2/n$. The CLT gives us

$$\hat{\pi}_1 \overset{\bullet}{\sim} N\left(\pi_1, \frac{\pi_1(1-\pi_1)}{n_1}\right) \perp \hat{\pi}_2 \overset{\bullet}{\sim} N\left(\pi_2, \frac{\pi_2(1-\pi_2)}{n_2}\right),$$

and so

$$\hat{\pi}_1 - \hat{\pi}_2 \overset{\bullet}{\sim} N\left(\pi_1 - \pi_2, \frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}\right).$$

Under $H_0 : \pi_1 = \pi_2$ and $n_1 = n_2$ the test statistic is

$$Z = \frac{\hat{\pi}_1 - \hat{\pi}_2}{\sqrt{2\hat{\pi}(1-\hat{\pi})/n}},$$

where $\hat{\pi} = (Y_1 + Y_2)/(2n)$ is pooled estimator, i.e. MLE under H_0.

Testing $H_0 : \pi_1 = \pi_2$ from two samples

Similar computations as in the one-sample case leads to

$$n_1 = n_2 = (z_{\alpha/2} + z_\beta)^2 \frac{\pi_1(1-\pi_1) + \pi_2(1-\pi_2)}{(\pi_1 - \pi_2)^2}.$$

Note that for $\alpha = 0.05$ and $\beta = 0.1$ we have $z_{0.025} = 1.960$ and $z_{0.1} = 1.282$. $1 - \beta = 0.99$ yields $z_{0.01} = 2.326$.
What happens when $\pi_1 \approx \pi_2$?

6.6.2 Sample size for simple logistic regression*

Let

$$\text{logit } \pi(x) = \alpha + \beta X,$$

where $X \sim N(\mu, \sigma^2)$ and

$$\tau = \log\left\{\frac{\pi(\mu+\sigma)/[1-\pi(\mu+\sigma)]}{\pi(\mu)/[1-\pi(\mu)]}\right\},$$

the log of the ratio of event odds when $x = \mu + \sigma$ and $x = \mu$. Then to test $H_0 : \beta \leq 0$ versus $H_0 : \beta > 0$ (or the other direction) at significance α and power $1 - \beta$ we need sample size

$$n = [z_\alpha + z_\beta e^{-\tau^2/4}]^2[1 + 2\pi(\mu)\delta]/[\pi(\mu)\tau^2],$$

where

$$\delta = [1 + (1+\tau^2)e^{5\tau^2/4}]/[1 + e^{-\tau^2/4}].$$

Cholesterol and heart disease, p. 239

- X is cholesterol level, Y indicates "severe heart disease."
- Know $\pi(\mu) = 0.08$. Want to be able to detect a 50% increase in probability for a standard deviation increase in cholesterol. 50% increase in probability is $1.5 \times 0.08 = 0.12$.
- $\pi(\mu)/[1 - \pi(\mu)] = 0.08/0.92 = 0.087$.
- $\pi(\mu + \sigma)/[1 - \pi(\mu + \sigma)] = 0.12/0.88 = 0.136$. So the odds ratio is $0.136/0.087 = 1.57$, and $\tau = \log(1.57) = 0.45$.
- Then for $\alpha = 0.05$, $1 - \beta = 0.9$, we have $\delta = 1.306$ and $n = 612$.
- Note: didn't need to know μ and σ, but rather $\pi(\mu)$ and $\pi(\mu + \sigma)$.

6.5.3 Sample size for one effect in multiple logistic regression*

Say now that we're interested in X_1 but there's $p-2$ more more predictors $X_2, \ldots, X_{p-1}$. Let R denote the multiple correlation between X and the remaining predictors:

$$R = \max_{||\mathbf{a}||=1} \{\text{corr}(X_1, a_2X_2 + \cdots + a_{p-1}X_{p-1})\}.$$

Let $\pi(\boldsymbol{\mu}) = \pi(\mu_1, \mu_2, \ldots, \mu_{p-1})$ be the probability at the mean of all $p-1$ variables.

τ is the now the log odds ratio comparing $\pi(\mu_1 + \sigma_1, \mu_2, \ldots, \mu_{p-1})$ to $\pi(\mu_1, \mu_2, \ldots, \mu_{p-1})$.

$$n = [z_\alpha + z_\beta e^{-\tau^2/4}]^2[1 + 2\pi(\boldsymbol{\mu})\delta]/[\pi(\boldsymbol{\mu})\tau^2(1-R^2)].$$

Heart disease example (continued):

- Say we have another variable X_2 is blood pressure and $R = \text{corr}(X_1, X_2) = 0.4$.
- Then $n = 612/(1 - 0.4^2) = 729$.
- What happens when $\text{corr}(X_1, X_2) \approx 1$. Is this problematic? Hint: think about the interpretation of β_1.

6.6.4, 6.6.5, & 6.6.6 Misc. power and sample size considerations

Read over if interested.