7. Cross tabulation and logistic regression

(PROC FREQ and PROC LOGISTIC)

Use of SAS
March 2010

Table analyses: PROC FREQ

```
/* One-way table */
proc freq data=afrika.bissau;
  tables dead;
run;

/* Two-way table */
proc freq data=afrika.bissau;
  tables bcg*dead;
run;

/* Three-way table */
proc freq data=afrika.bissau;
  tables agemm*bcg*dead;
run;
```

etc.

Two-way table (2 x 2 table)

```
proc freq data=afrika.bissau;
  tables bcg*dead;
run;
```

Table of bcg by dead

bcg	dead		
Frequency			
Percent			
Row Pct			
Col Pct	1	2	Total
1	124	3176	3300
	2.35	60.23	62.58
	3.76	96.24	
	56.11	62.87	
2	97	1876	1973
	1.84	35.58	37.42
	4.92	95.08	
	43.89	37.13	
Total	221	5052	5273
	4.19	95.81	100.00

Two-way table (2 x 2 table)

	Outcome		
Exposure	Yes	No	Total
Yes	a	b	n_1
No	c	d	n_2
Total	$a+c$	$b+d$	n

Hypothesis H_0: The probability of having the outcome is the same in the two exposure groups.

The probability of having the outcome under H_0:

$$p = \frac{a+c}{n}$$

.

The EXPECTED numbers under H_0 in the four cells are calculated as:

	Outcome		
Exposure	Yes	No	Total
Yes	$E(a) = p \times n_1$	$E(b) = (1-p) \times n_1$	n_1
No	$E(c) = p \times n_2$	$E(d) = (1-p) \times n_2$	n_2
Total	$a + c$	$b + d$	n

Chi-square test for testing H_0 (observed - expected):

$$\begin{aligned} X^2 &= \frac{(a - E(a))^2}{E(a)} + \frac{(b - E(b))^2}{E(b)} + \frac{(c - E(c))^2}{E(c)} + \frac{(d - E(d))^2}{E(d)} \\ &\sim \chi^2(1) \end{aligned}$$

H_0 is rejected if p-value < 0.05 which corresponds to $X^2 > 3.84$.

Risk of Dying and BCG

```
proc freq data=afrika.bissau;
  tables bcg*dead / expected chisq nocol nopercent;
run;
```

Table of bcg by dead

bcg	dead		
Frequency|			
Expected |			
Row Pct |	1	2	Total
1	124	3176	3300
	138.31	3161.7	
	3.76	96.24	
2	97	1876	1973
	82.692	1890.3	
	4.92	95.08	
Total	221	5052	5273

Statistics for Table of bcg by dead

Statistic	DF	Value	Prob
Chi-Square	1	4.1291	0.0422
Likelihood Ratio Chi-Square	1	4.0516	0.0441
Continuity Adj. Chi-Square	1	3.8456	0.0499
Mantel-Haenszel Chi-Square	1	4.1283	0.0422
Phi Coefficient		-0.0280	
Contingency Coefficient		0.0280	
Cramer's V		-0.0280	

Risk Ratio

	Outcome		
Exposure	Yes	No	Total
Yes	a	b	n_1
No	c	d	n_2
Total	$a+c$	$b+d$	n

Risk ratio:

$$\text{RR} = \frac{\text{probability of outcome among exposed}}{\text{probability of outcome among not-exposed}} = \frac{a/n_1}{c/n_2}.$$

The H_0 corresponds to RR $= 1$.

Odds ratio

	Outcome		
Exposure	Yes	No	Total
Yes	a	b	n_1
No	c	d	n_2
Total	$a+c$	$b+d$	n

Odds ratio:

$$\text{OR} = \frac{\text{odds of outcome among exposed}}{\text{odds of outcome among not-exposed}} = \frac{a/b}{c/d} = \frac{a \times d}{b \times c}$$

The H_0 corresponds to OR $= 1$.

RR and OR in `proc freq`

```
proc freq data=afrika.bissau;
  table bcg*dead / RELRISK nocol nopercent;
run;
```

Table of bcg by dead

bcg	dead		
Frequency|			
Row Pct |	1	2	Total
1	124	3176	3300
	3.76	96.24	
2	97	1876	1973
	4.92	95.08	
Total	221	5052	5273

Statistics for Table of bcg by dead

Estimates of the Relative Risk (Row1/Row2)

Type of Study	Value	95% Confidence Limits	
Case-Control (Odds Ratio)	0.7551	0.5754	0.9909
Cohort (Col1 Risk)	0.7643	0.5895	0.9910
Cohort (Col2 Risk)	1.0122	1.0000	1.0245

OR and RR in `proc freq`

It is important that the two variables in the `table` statement are coded properly when using the OR and RR:

```
data hope;
  set afrika.bissau;
  if dead=2 then deadny=0;
  if dead=1 then deadny=1;
proc freq data=hope;
  table bcg*deadny / relrisk nocol nopercent norow;
run;
```

```
                              bcg       deadny

                              Frequency|
                              Row Pct  |       0|       1|  Total
                              ---------+--------+--------+
                                     1 |   3176 |    124 |   3300
                              ---------+--------+--------+
                                     2 |   1876 |     97 |   1973
                              ---------+--------+--------+
                              Total        5052      221     5273
```

```
                    Statistics for Table of bcg by deadny
                 Estimates of the Relative Risk (Row1/Row2)
Type of Study                    Value       95% Confidence Limits
-----------------------------------------------------------------------
Case-Control (Odds Ratio)       1.3243        1.0092        1.7378
Cohort (Col1 Risk)              1.0122        1.0000        1.0245
Cohort (Col2 Risk)              0.7643        0.5895        0.9910
```

R x C tables

```
proc freq data=afrika.bissau;
  table ethnic*dead / chisq;
run;
```

ethnic	dead		
Frequency Percent Row Pct Col Pct	1	2	Total
Balanta	37 0.70 4.48 16.74	788 14.94 95.52 15.60	825 15.65
Fula	52 0.99 3.66 23.53	1370 25.98 96.34 27.12	1422 26.97
Mandinga	49 0.93 4.22 22.17	1113 21.11 95.78 22.03	1162 22.04
Other	23 0.44 3.08 10.41	724 13.73 96.92 14.33	747 14.17
Pepel	60 1.14 5.37 27.15	1057 20.05 94.63 20.92	1117 21.18
Total	221 4.19	5052 95.81	5273 100.00

Statistics for Table of ethnic by dead

Statistic	DF	Value	Prob
Chi-Square	4	7.3670	0.1177
Likelihood Ratio Chi-Square	4	7.3268	0.1196
Mantel-Haenszel Chi-Square	1	1.0857	0.2974
Phi Coefficient		0.0374	
Contingency Coefficient		0.0374	
Cramer's V		0.0374	

Exercise: PROC FREQ

Using the bissau data:

1. Investigate whether DTP-vaccinated children (variable `dtp`) dies more often than DTP-unvaccinated children.
2. Calculate the odds ratio (OR) and corresponding 95% confidence interval.
3. The variable `region` indicates the rural region of the children. Is mortality associated with region?

Logistic regression: PROC LOGISTIC

Logistic regression is like a linear regression, but here the outcome is DISCRETE with two levels (yes/no, died/survived, ill/well). Look again at the 2 x 2 table

	Outcome		
Exposure	Yes	No	Total
Yes	a	b	n_1
No	c	d	n_2

Let $p = a/n_1$ be the probability of outcome among exposed. Odds can then be defined as

$$\text{odds} = \frac{p}{1-p} = \frac{a/n_1}{1-a/n_1} = \frac{a/n_1}{b/n_1} = \frac{a}{b}$$

Logistic regression for 2 x 2 table

What is modeled in a logistic regression is the NATURAL LOGARITHM of the ODDS of outcome:

$$\ln(\text{odds}) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X,$$

where X is the exposure covariate. We call ln(odds) for the LOG-ODDS. Now let us assume that the exposure is coded like

$$X = \begin{cases} 1 & \text{Exposed} \\ 0 & \text{Non-exposed} \end{cases}$$

The log-odds of outcome among exposed ($X = 1$) is

$$\ln\left(\frac{p_1}{1-p_1}\right) = \beta_0 + \beta_1 \times 1 = \beta_0 + \beta_1.$$

The log-odds of outcome among non-exposed ($X = 0$) is

$$\ln\left(\frac{p_0}{1-p_0}\right) = \beta_0 + \beta_1 \times 0 = \beta_0.$$

The β_0 is the log-odds of outcome in non-exposed.

Now, the difference in log-odds between exposed and non-exposed is

$$\ln\left(\frac{p_1}{1-p_1}\right) - \ln\left(\frac{p_0}{1-p_0}\right) = \beta_0 + \beta_1 - \beta_0 = \beta_1$$

Using the rule of logarithms

$$\ln(a) - \ln(b) = \ln(\frac{a}{b})$$

we get

$$\ln\left(\frac{p_1/(1-p_1)}{p_0/(1-p_0)}\right) = \beta_1$$

and

$$\frac{p_1/(1-p_1)}{p_0/(1-p_0)} = \exp(\beta_1)$$

This means that the odds ratio between exposed and non-exposed is

$$\text{OR} = \exp(\beta_1).$$

Estimation of the regression coefficients is done using maximum likelihood.

PROC LOGISTIC

```
proc logistic data=afrika.bissau;
  class bcg / param=ref;
  model dead(event="1")=bcg;
run;
```

REMEMBER the option `param=ref`

Part of output:

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-2.9621	0.1041	809.3011	<.0001
bcg	1	1	-0.2810	0.1386	4.1074	0.0427

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
bcg 1 vs 2	0.755	0.575	0.991

Multiple logistic regression

$$\ln(\text{odds}) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots ,$$

The interpretation is still that $\exp(\beta_1)$ is an odds ratios, but now adjusted for the covariates $X_2, X_3, \cdots$. The same idea as in linear regression.

Remember that the response or outcome is discrete with two categories. However the covariates $(X_1, X_2, X_3, \cdots)$ do not need to be categorical, they can also be continuous.

In SAS one use the CLASS statement to indicate categorical variables, and variables in a MODEL statement that is not listed in the CLASS statement is assumed to be continuous.

Multiple logistic regression: PROC LOGISTIC

```
proc logistic data=afrika.bissau;
  class bcg / param=ref;
  model dead(event="1")=bcg agemm;
run;
```

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
bcg	1	5.4366	0.0197
agemm	1	1.5307	0.2160

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-3.0510	0.1281	567.4989	<.0001
bcg	1	1	-0.3450	0.1480	5.4366	0.0197
agemm		1	0.0486	0.0393	1.5307	0.2160

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
bcg 1 vs 2	0.708	0.530	0.946
agemm	1.050	0.972	1.134

Interpretation: For each increase of 1 in `agemm` the odds increases with 1.050.

Multiple logistic regression: PROC LOGISTIC

The variable agemm is now used as a CLASS variable:

```
proc logistic data=afrika.bissau;
  class bcg agemm / param=ref;
  model dead(event="1")=bcg agemm;
run;
```

agemm has 7 classes: 0 to 6. SAS automatically generates 7 indicator functions for each class and include 6 of these in the regression model. The class not included (SAS uses per default the highest class) is the reference against which all other classes are listed in the output.

The test in TYPE 3 for agemm is a test for the hypothesis of equal risk of dying in the 7 classes. This test does not change when you change reference.

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
bcg	1	5.2393	0.0221
agemm	6	7.3938	0.2860

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-3.0948	0.3281	88.9498	<.0001
bcg	1	1	-0.3430	0.1499	5.2393	0.0221
agemm	0	1	0.0588	0.3599	0.0267	0.8703
agemm	1	1	0.1803	0.3513	0.2635	0.6077
agemm	2	1	-0.1925	0.3650	0.2783	0.5978
agemm	3	1	0.2700	0.3514	0.5904	0.4423
agemm	4	1	0.4044	0.3492	1.3410	0.2469
agemm	5	1	0.3618	0.3549	1.0392	0.3080

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
bcg 1 vs 2	0.710	0.529	0.952
agemm 0 vs 6	1.061	0.524	2.147
agemm 1 vs 6	1.198	0.602	2.384
agemm 2 vs 6	0.825	0.403	1.687
agemm 3 vs 6	1.310	0.658	2.608
agemm 4 vs 6	1.498	0.756	2.971
agemm 5 vs 6	1.436	0.716	2.879

Change of reference group: REF=""

The variable `agemm` is again used as a CLASS variable but now choosing agegroup 4 as reference:

```
proc logistic data=afrika.bissau;
  class bcg agemm(ref="4") / param=ref;
  model dead(event="1")=bcg agemm;
run;
```

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
bcg	1	5.2393	0.0221
agemm	6	7.3938	0.2860

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-2.6904	0.1973	185.9085	<.0001
bcg	1	1	-0.3430	0.1499	5.2393	0.0221
agemm	0	1	-0.3456	0.2479	1.9444	0.1632
agemm	1	1	-0.2241	0.2368	0.8953	0.3441
agemm	2	1	-0.5969	0.2575	5.3731	0.0204
agemm	3	1	-0.1344	0.2385	0.3175	0.5731
agemm	5	1	-0.0426	0.2442	0.0304	0.8616
agemm	6	1	-0.4044	0.3492	1.3410	0.2469

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
bcg 1 vs 2	0.710	0.529	0.952
agemm 0 vs 4	0.708	0.435	1.150
agemm 1 vs 4	0.799	0.502	1.271
agemm 2 vs 4	0.551	0.332	0.912
agemm 3 vs 4	0.874	0.548	1.395
agemm 5 vs 4	0.958	0.594	1.546
agemm 6 vs 4	0.667	0.337	1.323

Exercise: PROC LOGISTIC

Using the Bissau data:

1. Make a logistic regression where outcome is `dead` and exposure is `dtp`. Interpret the results and compare with the results from the exercise using `proc freq` on page 14.
2. Now control for `bcg` in the logistic regression from 1 above. What happend with the odds ratio for `dtp`?
3. Add variables `agemm` and `region` to the model as class variables. Let `region=7` be the reference group for variable `region`. Did inclusion of these variables change interpretation of effect `dtp`?

Logistic regression: PROC GENMOD

The procedure `proc genmod` (GENeralized linear MODels) can also perform logistic regression:

```
proc genmod data=afrika.bissau;
  class bcg;
  model dead=bcg / dist=binomial;
  estimate "BCG+ vs BCG-" bcg 1 -1 / exp;
run;
```

Getting the odds ratios and their confidence intervals out is however not as easy as in `proc logistic`, but done with `estimate` statements. `proc genmod` is however very flexible and can analyse other types of regression models like Poisson (event per person-years) and linear regression.