7. Categorical data

Karl B Christensen
http://publicifsv.sund.ku.dk/~kach/SPSS

Analysis of categorical data

- Tables (frequency tables)
- Risk ratios (relative risks)
- Odds ratios
- Logistic regression

Table

Exposure	Outcome: Yes	Outcome: 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 Guinea-Bissau data set

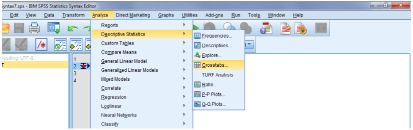
The data set called `bissau.sav` comes from rural Guinea-Bissau, West-Africa: 5273 children visited at age < 7 months and followed for approximately six months. Registration of vaccination status and deaths registered during follow-up.

```
GET FILE='P:\bissau.sav'.
```

Crosstabs

```
CROSSTABS
  /TABLES=bcg BY dead
  /FORMAT=AVALUE TABLES
  /CELLS=COUNT ROW
  /COUNT ROUND CELL.
```

Note that `ROW` gives us row-percentages



Crosstabs

Note that `ROW` gives us row-percentages: risk of dying 3.8% and 4.9%, repsectively

bcg * dead Crosstabulation

			dead		
			1	2	Total
bcg	1	Count	124	3176	3300
		% within bcg	3,8%	96,2%	100,0%
	2	Count	97	1876	1973
		% within bcg	4,9%	95,1%	100,0%
Total		Count	221	5052	5273
		% within bcg	4,2%	95,8%	100,0%

Crosstabs

- The risk of dying in the two BCG groups: 3.8% with BCG and 4.9% without BCG.
- We want to know if these probabilities are significantly different.
- We test the null hypothesis

  H_0: the probability of dying is the same in the two groups.

Observed 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: probability of outcome is the same in the two exposure groups.

probability of outcome under H_0 is $p = \frac{a+c}{n}$.

Expected table

Under H_0 expected numbers in the four cells are:

	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):

$$X^2 = \frac{[a - E(a)]^2}{E(a)} + \frac{[b - E(b)]^2}{E(b)} + \frac{[c - E(c)]^2}{E(c)} + \frac{[d - E(d)]^2}{E(d)}$$

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

Expected table

```
CROSSTABS
  /TABLES=bcg BY dead
  /FORMAT=AVALUE TABLES
  /CELLS=COUNT ROW EXPECTED
  /COUNT ROUND CELL.
```

bcg * dead Crosstabulation

			dead		Total
			1	2	
bcg	1	Count	124	3176	3300
		Expected Count	138,3	3161,7	3300,0
		% within bcg	3,8%	96,2%	100,0%
	2	Count	97	1876	1973
		Expected Count	82,7	1890,3	1973,0
		% within bcg	4,9%	95,1%	100,0%
Total		Count	221	5052	5273
		Expected Count	221,0	5052,0	5273,0
		% within bcg	4,2%	95,8%	100,0%

Risk of Dying and BCG - Chi-square test

Compute

$$X^2 = \frac{[a - E(a)]^2}{E(a)} + \frac{[b - E(b)]^2}{E(b)} + \frac{[c - E(c)]^2}{E(c)} + \frac{[d - E(d)]^2}{E(d)}$$

i.e.

$$X^2 = \frac{[124 - 138.3]^2}{138.3} + .. + \frac{[1876 - 1890.3]^2}{1890.3} = ..$$

or

```
CROSSTABS
  /TABLES=bcg BY dead
  /FORMAT=AVALUE TABLES
  /STATISTICS=CHISQ
  /CELLS=COUNT ROW
  /COUNT ROUND CELL.
```

Risk of Dying and BCG - Chi-square test

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	4,129[a]	1	,042		
Continuity Correction[b]	3,846	1	,050		
Likelihood Ratio	4,052	1	,044		
Fisher's Exact Test				,047	,026
Linear-by-Linear Association	4,128	1	,042		
N of Valid Cases	5273				

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 82,69.

b. Computed only for a 2x2 table

Risk of Dying and BCG - Chi-square test

The risk of dying in the two BCG groups: 3.76% and 4.92%.

- We see from the Chi-square test that the probability of dying differs significantly between the groups.
- How can we quantify this?
  - Risk difference

$$4.92 - 3.76 = 1.16$$

  - Relative risk

$$\frac{4.92}{3.76} = 1.31 \text{ or } \frac{3.76}{4.92} = \frac{1}{1.31}$$

- SPSS will not estimate the risk difference.
- If you want SPSS to calculate relative risk

  SPSS 'assumes' that the reference group is the first row and the outcome of interest in the first column.

Risk Ratio

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

Risk ratio:

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

The H_0 corresponds to $\mathrm{RR} = 1$.

Odds

Exposure	Outcome 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

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

does not contain any other information than the probability. If the probability is higher odds are higher and vice versa.

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 $\text{OR} = 1$.

```
/TABLES=bcg BY dead
/FORMAT=AVALUE TABLES
/STATISTICS=RISK
/CELLS=COUNT ROW
/COUNT ROUND CELL.
```

output

Risk Estimate

	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for bcg (1 / 2)	,755	,575	,991
For cohort dead = 1	,764	,589	,991
For cohort dead = 2	1,012	1,000	1,024
N of Valid Cases	5273		

Exercise using the bissau data 1

1. Do DTP-vaccinated children (variable `dtp`) die more often than DTP-unvaccinated children?
2. Calculate the odds ratio (OR) and corresponding 95% confidence interval.

R x C tables

We can also compare more than two groups

```
/TABLES=region BY dead
/FORMAT=AVALUE TABLES
/STATISTICS=CHISQ
/CELLS=COUNT ROW
/COUNT ROUND CELL.
```

The null hypothesis

$$H_0\text{: the risk of dying is the same in the five groups}$$

Null hypothesis H_0: risk of dying is the same in the five groups

region * dead Crosstabulation

			dead 1	dead 2	Total
region	1	Count	50	1065	1115
		% within region	4,5%	95,5%	100,0%
	2	Count	69	1246	1315
		% within region	5,2%	94,8%	100,0%
	5	Count	44	1041	1085
		% within region	4,1%	95,9%	100,0%
	7	Count	24	771	795
		% within region	3,0%	97,0%	100,0%
	8	Count	34	929	963
		% within region	3,5%	96,5%	100,0%
Total		Count	221	5052	5273
		% within region	4,2%	95,8%	100,0%

Null hypothesis H_0: risk of dying is the same in the five groups

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	7,707[a]	4	,103
Likelihood Ratio	7,782	4	,100
Linear-by-Linear Association	5,514	1	,019
N of Valid Cases	5273		

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 33,32.

The variable `ethnic` indicates the ethnic group the child belongs to.

1. Is mortality associated with this variable?

Logistic regression

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

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

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

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) the log-odds. 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 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 this means that the odds ratio between exposed and non-exposed is

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

Estimation of the regression coefficients is done using maximum likelihood.

```
LOGISTIC REGRESSION VARIABLES dead
  /METHOD=ENTER bcg
  /CONTRAST (bcg)=Indicator
  /PRINT=CI(95)
  /CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20) CUT(0.5).
```

request confidence interval for OR

http://publicifsv.sund.ku.dk/~kach/SPSS/F7_gif1.gif

For the case of a 2×2 table the logistic regression model is just a more complicated way of getting the OR with a general way of writing the model

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

the exposure covariate X was coded

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

Logistic regression exercise

Using the Bissau data

1. Make a logistic regression where outcome is `dead` and exposure is `dtp`.
2. Interpret the results and compare with the results from the exercise using `CROSSTABS`.

Logistic regression

For the case of a 2×2 table the logistic regression model is just a more complicated way of getting the OR with a general way of writing the model

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

the exposure covariate X was coded

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

this general framework also works for linear effect of X (e.g. age).

Linear effect of covariate

- The response or outcome is discrete with two categories.
- Covariates $(X_1, X_2, X_3, \cdots)$: The effect of the X's can be modelled as a linear effect (comparing risk for $X = x$ to risk for $X = x + 1$)
- Indicate categorical X's

Multiple logistic regression

The response (or outcome) is discrete with two categories.

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

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

Logistic regression: coding of the outcome variable

SPSS models the probability

$$P(Y = 1)$$

if $Y \in \{0, 1\}$.
But what about `dead` ?
Recode

$$\texttt{dead2} = \begin{cases} 1, & \texttt{dead} = 1 \\ 0, & \texttt{dead} = 2 \end{cases}$$

the effect of bcg adjusted for the (linear) effect of age

```
LOGISTIC REGRESSION VARIABLES dead2
  /METHOD=ENTER agemm bcg
  /CONTRAST (bcg)=Indicator
  /CRITERIA=PIN(.05) POUT(.10) ITERATE(20) CUT(.5).
```

output

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	bcg(1)	,345	,148	5,433	1	,020	1,412	1,056	1,887
	agemm	-,049	,039	1,530	1	,216	,953	,882	1,029
	Constant	3,051	,128	567,488	1	,000	21,138		

a. Variable(s) entered on step 1: bcg, agemm.

Multiple logistic regression exercise

Using the Bissau data

1. Make a logistic regression where outcome is `dead` and exposure is `dtp` and `agemm`. Interpret the parameters.
2. Now control for `bcg` in the logistic regression model. What happened to the effect of `dtp` ?