1 Contingency Tables

1.1 Basics

A contingency table is a place where a joint distribution of two categorical variables, say X and Y, where X has I levels and Y has J levels, can be summarized. It is usual to consider an X variable to be an explanatory variable, and a Y variable a response variable. Both marginal distributions and conditional distributions can be of interest. Data that is used can either be offered in raw form (Format 1) or in summarized form (Format 2) in SPSS.

Example 1
The file HockeyGenderRaw contains two variables. The response variable WatY (Yes-1, No-2) records whether or not an individual watches hockey and appears in column 1 and the explanatory variable GenX (Male -1, Female-2) records gender and appears in column 2.
The file WatchHockeyGenderSummarized summarizes the counts of observations for the (GenX, WatY) pairs (1,1), (1,2), (2,1) and (2,2).

Format 1

WatY	GenX
1	1
1	1
⋮	⋮
1	2
1	2
⋮	⋮
2	1
2	1
⋮	⋮
2	2
2	2
⋮	⋮

Format 2

WatY	GenX	Count
1	1	33
1	2	9
2	1	11
2	2	36

When working with raw data with columns of categorical variables, one should create a summary cross-tab table that counts the groupings. Here are the SPSS commands and output to do so for the hockey watching and gender raw data.

1. Analyze>Descriptive Statistics>Crosstabs
2. Row(s): GenX
3. Column(s): WatY
4. OK

GenX * WatY Crosstabulation

			WatY		Total
			1	2	
GenX	1	Count	33	11	44
		% within GenX	75.0%	25.0%	100.0%
	2	Count	9	36	45
		% within GenX	20.0%	80.0%	100.0%
Total		Count	42	47	89
		% within GenX	47.2%	52.8%	100.0%

1.2 Comparing two categorical variables(each with 2 levels)

It is usual to have X in the row, and Y in the column in a contingency table. Here n_{ij} represents the count of observations in the ij-th cell, $i = 1, 2$ and $j = 1, 2$, while n_{i+} is the count in row i, n_{+j} is the count in column j, and n is the total number of observations.

		Y		
		1	2	Total
X	1	n_{11}	n_{12}	n_{1+}
	2	n_{21}	n_{22}	n_{2+}
	Total	n_{+1}	n_{+2}	n

Conditional Probabilities:
$P(Y = 1|X = 1) = \pi_1$ with $\hat{\pi}_1 = p_1 = n_{11}/n_{1+}$
$P(Y = 1|X = 2) = \pi_2$ with $\hat{\pi}_2 = p_2 = n_{21}/n_{2+}$

- Difference in probabilities π_1 and π_2:

  Wald $(1-\alpha) \times 100\%$ confidence interval for $\pi_1 - \pi_2$: $(p_1 - p_2) \pm z_{\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$

- Relative Risk $rr = \pi_1/\pi_2$, $\hat{rr} = p_1/p_2$:

  $(1-\alpha) \times 100\%$ confidence interval for $\ln(rr)$: $\ln(p_1/p_2) \pm z_{\alpha/2}\sqrt{\frac{1-p_1}{n_1 p_1} + \frac{1-p_1}{n_1 p_1}}$

  Convert (exponentiate) the endpoints to find a CI for rr.

- Odds Ratio
  Odds of success(=1) in row $i = \pi_i/(1-\pi_i)$
  Odds Ratio= $\vartheta = odds_1/odds_2$

  MLE for ϑ is $\hat{\vartheta} = n_{11}n_{22}/n_{12}n_{21}$

  $(1-\alpha) \times 100\%$ confidence interval for $\ln(\vartheta)$: $\ln(\hat{\vartheta}) \pm z_{\alpha/2}\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$

  Convert (exponentiate) the endpoints to find a CI for odds ratio.

Example 2
From class

PCR	Relapse	Count
1	1	30
1	2	45
2	1	8
2	2	95

PCR can be considered to the explanatory variable, and Relapse the response

SPSS commands and output:

1. Data>Weight Cases
   Weight by Frequency Variable: Total
   OK

2. Analysis>Descriptive Statistics>Crosstabs
   Row(s): PCR
   Column(s): Relapse
   Statistics popup: check Risk
   OK
   Cells popup: Counts: check Observed
   Percentages: check Row
   OK
   OK

*Note: If data is in two columns, one for PCR (X) and one for Relapse (Y), the weight cases command lines are unnecessary.

PCR * Relapse Crosstabulation

			Relapse		
			1	2	Total
PCR	1	Count	30	45	75
		% within PCR	40.0%	60.0%	100.0%
	2	Count	8	95	103
		% within PCR	7.8%	92.2%	100.0%
Total		Count	38	140	178
		% within PCR	21.3%	78.7%	100.0%

Risk Estimate

	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for PCR (1 / 2)	7.917	3.361	18.648
For cohort Relapse = 1	5.150	2.504	10.590
For cohort Relapse = 2	.651	.536	.789
N of Valid Cases	178		

$p_1 = 0.4, p_2 = 0.078$ 95% Wald CI for $(\pi_1 - \pi_2)$ calculate by hand: (.200, 0.445), or use the app at `http://vassarstats.net/prop2_ind.html`, or

1. Analyze>Generalized Linear Models>Generalized Linear Models

2. In "Type of Model" tab:

   > Custom, and "Distribution": Binomial

3. In "Response" tab:

   Choose dependent variable: Response

4. In "Predictors" tab:

   Choose factor: pcr

5. In "Model" tab:

   Move pcr into "Model"

6. Click OK

7. The confidence interval can be found in the "Parameter Estimates" table in row labelled pcr=1.00

With 95% confidence we learn that the proportion of PCR positive children who relapse is between 0.2 and 0.445 larger than the proportion of PCR negative children who relapse.
Note that 0 is not in the 95% CI, indicating a significant difference in the relapse rate for PCR positive and negative children.

Estimated **relative risk** $\hat{rr} = p_1/p_2 = 0.4/0.078 = 5.15$,
95% CI for rr of a relapse of PCR positive versus PCR negative children falls between (2.504, 10.590)

[from output). With 95% confidence we find that the probability of a relapse is between 2.5 and 10.6 times higher for children with a positive PCR than a negative PCR.
Note that 1 is not in the 95% CI, we have sufficient evidence that the probability of a relapse is significantly higher for PCR positive children.

Estimated **odds ratio** $= \hat{\vartheta} = n_{11}n_{22}/n_{12}n_{21} = (30\text{x}95)/(8\text{x}45) = 7.917$
95% CI for odds ratio of a relapse for PCR positive versus PCR negative children falls between (3.361, 18.648) [from output].
With 95% confidence we find that the odds of a relapse is between 3.4 and 18.6 times higher for children with a positive PCR than a negative PCR.
Note that 1 is not in the 95% CI, we have significant evidence that a relapse is associated with PCR. The odds of a relapse is 3.361 to 18.648 times higher for the PCR+ group than for the PCR- group.

Mantel-Haenzel Odds Ratio Test

SPSS commands and output:

1. Analysis>Descriptive Statistics>Crosstabs
   Row(s): PCR
   Column(s): Relapse
   Statistics popup: check Cochran's and Mantel-Haenszel statistics
   check Test common odds ratio equals: 1
   OK
   Cells popup: Counts: check Observed
   Percentages: check Row
   OK
   OK

Mantel-Haenszel Common Odds Ratio Estimate

Estimate			7.917
ln(Estimate)			2.069
Std. Error of ln(Estimate)			.437
Asymp. Sig. (2-sided)			.000
Asymp. 95% Confidence Interval	Common Odds Ratio	Lower Bound	3.361
		Upper Bound	18.648
	ln(Common Odds Ratio)	Lower Bound	1.212
		Upper Bound	2.926

The Mantel-Haenszel common odds ratio estimate is asymptotically normally distributed under the common odds ratio of 1.000 assumption. So is the natural log of the estimate.

1. $H_0 : \vartheta = 1$ versus $H_a : \vartheta \neq 1$, $\alpha = 0.05$
2. Random sample (ok), large enough sample (each cell has at least count 5)
3. $z_0 = 2.069/.437 = 4.73$
4. P-value<0.001
5. P-value$< \alpha = 0.05$, reject H_0
6. At significance level of 5% the data provide sufficient evidence that the odds ratio of a relapse for PCR positive and negative children is different from 1, indicating a relationship between PCR outcome and relapse.