Biostatistics 103: Qualitative Data – Tests of Independence

Y H Chan

Parametric & non-parametric tests[1] are used when the outcome response is quantitative and our interest is to determine whether there are any statistical differences between/amongst groups (which are categorical).

In this article, we are going to discuss how to analyse relationships between categorical variables. Table I shows the first five cases of 200 subjects with their gender and intensity of snoring (No, At Times, Frequent and Always) and snoring status (Yes or No) recorded.

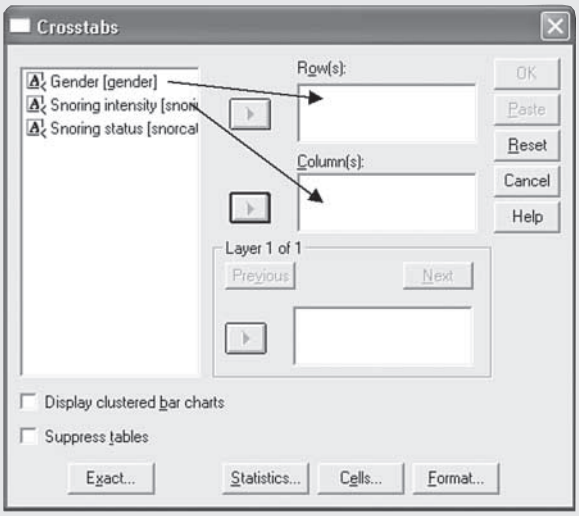
Table I. Data structure in SPSS.

Subject	Gender	Snoring Intensity	Snoring Status
1	Male	No	No
2	Male	Always	Yes
3	Female	Frequent	Yes
4	Male	At times	Yes
5	Female	No	No

Here, we have two interests. One is to determine whether there's an association between gender and snoring intensity and the other is the association between gender and snoring status. The interpretation of the results for both analyses is not similar.

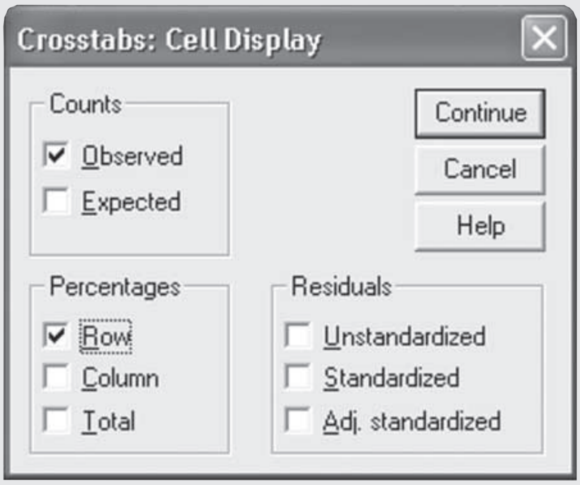
Let's discuss the 1st interest. The null hypothesis is: **There is *No Association* between gender and snoring intensity.** To test this hypothesis of no association (or **independence**), the **Chi-Square test** is performed. With the given data structure in Table I, to perform the Chi-Square test in SPSS, use ***Analyse, Descriptive Statistics***, ***Crosstabs*** and the following template is obtained:

Template I. Crosstabs.



It does not matter whether we put Snoring intensity or Gender into the **Row(s)** or **Columns** but for "easier interpretation" of the results (later) it is recommended to put the "the categorical variable of outcome interest" (in this case, the Snoring intensity) in the **Columns** option. Click on the **Cells** button and tick the Row Percentages (the Observed Counts is ticked by default), then Continue.

Template II. Crosstabs: Cell Display.



Counts
☑ Observed
☐ Expected

Percentages
☑ Row
☐ Column
☐ Total

Residuals
☐ Unstandardized
☐ Standardized
☐ Adj. standardized

The crosstabulation table is shown in Table II. This table is a 2 X 4 (read as 2 by 4); 2 levels for Gender and 4 levels for Snoring intensity.

Clinical Trials and Epidemiology Research Unit
226 Outram Road
Blk A #02-02
Singapore 169039

Y H Chan, PhD
Head of Biostatistics

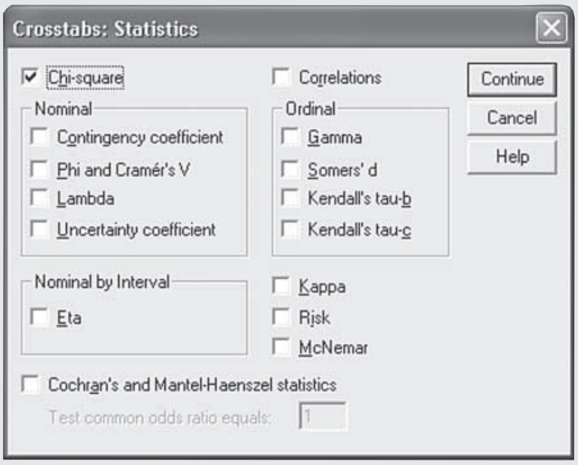
Correspondence to:
Y H Chan
Tel: (65) 6317 2121
Fax: (65) 6317 2122
Email: chanyh@cteru.com.sg

Table II. Crosstabulation table of Gender and Snoring intensity.

			Snoring intensity				
			ALWAYS	AT TIMES	FREQUENT	NO	Total
Gender	Female	Count	9	31	6	58	104
		% within Gender	8.7%	29.8%	5.8%	55.8%	100.0%
	Male	Count	23	26	6	41	96
		% within Gender	24.0%	27.1%	6.3%	42.7%	100.0%
Total		Count	32	57	12	99	200
		% within Gender	16.0%	28.5%	6.0%	49.5%	100.0%

To ask for the **Chi-Square test**, click on the **Statistics** button at the bottom of Template I and the Crosstabs:Statistics template is shown – tick the Chi-square box.

Template III. Crosstabs: Statistics.



Crosstabs: Statistics

☑ Chi-square
□ Correlations

Nominal
□ Contingency coefficient
□ Phi and Cramér's V
□ Lambda
□ Uncertainty coefficient

Ordinal
□ Gamma
□ Somers' d
□ Kendall's tau-b
□ Kendall's tau-c

Nominal by Interval
□ Eta

□ Kappa
□ Risk
□ McNemar

□ Cochran's and Mantel-Haenszel statistics
Test common odds ratio equals: 1

Table III gives the result for the Chi-Square test.

Table III. Chi-Square test result for the (2 X 4) Gender and Snoring intensity.

Chi-Square Tests			
	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	9.177[a]	3	.027
Continuity Correction			
Likelihood Ratio	9.390	3	.025
Linear-by-Linear Association	5.915	1	.015
N of Valid Cases	200		

[a] 0 cells (.0%) have expected count less than 5. The minimum expected count is 5.76.

Here the Pearson Chi-Square value is 9.17 with df (degree of freedom) = 3 and the p-value is 0.027 (<0.05) – the rest of the statistics in the table is of no interest to us! Hence we reject the null hypothesis of no association.

The Chi-Square test only tells us whether there is any association between two categorical variables but does not indicate what the association is. From Table II, by inspection, it is obvious that the difference lies in the males being more likely to have 'Always' snoring intensity compared to the females (24% vs 8.7%). Sometimes it's not so straightforward to interpret an association!

For the 2nd interest, the null hypothesis is: **There is *No Association* between gender and snoring status**. The (2 x 2) crosstabulation table and the Chi-Square test results are shown in tables IV and V respectively.

Table IV. (2 x 2) crosstabulation table of Gender and Snoring status.

Gender* Snoring status Crosstabulation					
			Snoring status		
			NO	YES	Total
Gender	Female	Count	58	46	104
		% within Gender	55.8%	44.2%	100.0%
	Male	Count	41	55	96
		% within Gender	42.7%	57.3%	100.0%
Total		Count	99	101	200
		% within Gender	49.5%	50.5%	100.0%

Table V. Result for Chi-Square test for the (2 X 2) Gender and Snoring status.

Chi-Square Tests					
	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	3.407[b]	1	.065		
Continuity Correction[a]	2.904	1	.088		
Likelihood Ratio	3.417	1	.065		
Fisher's Exact Test				.068	.044
Linear-by-Linear Association	3.390	1	.066		
N of Valid Cases	200				

[a] Computed only for a 2x2 table.

[b] 0 cells (.0%) have expected count less than five. The minimum expected count is 47.52.

This has be 0 for Pearson's Chi-Square to be valid

Here the Pearson Chi-Square p-value is 0.065 (>0.05) which means that there was no association between gender and snoring status. A different conclusion from the above results on the association between Gender and Snoring intensity!

You may have observed that the Chi-Square Tests Tables of III and V are different. The reason is that for a (2 x 2) association, SPSS automatically gives us the result for the **Fisher's Exact Test** whereas for a non (2 x 2), we have to "ask" for it (but we have to purchase this Exact test module). Why do we need this Fisher's Exact test?

The validity of the Pearson's Chi-Square test is violated when there are 'small frequencies' in the cells. The formal definitions of these assumptions (not reproduced here) for the validity can be found in any statistical textbook.

In SPSS, this validity is easily checked by observing the 'last line' of the Chi-Square Tests Table (for example in Table V), we want **0 cells (.0%) have expected count less than five**, otherwise we will have to use the Fisher's Exact test. Table VI and VII shows a situation where we should be cautious:

Table VI. 2 x 2 crosstabulation of Gender and Snoring status (n = 56)

Gender* Snoring status Crosstabulation

			Snoring status		
			NO	YES	Total
Gender	Female	Count	22	1	23
		% within Gender	95.7%	4.3%	100.0%
	Male	Count	25	8	33
		% within Gender	75.8%	24.2%	100.0
Total		Count	47	9	56
		% within Gender	83.9%	16.1%	100.0%

Table VII. Chi-Square test for table VI.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	3.977[b]	1	.046		
Continuity Correction[a]	2.693	1	.104		
Likelihood Ratio	4.594	1	.032		
Fisher's Exact Test				.067	.047
Linear-by-Linear Association	3.906	1	.048		
N of Valid Cases	56				

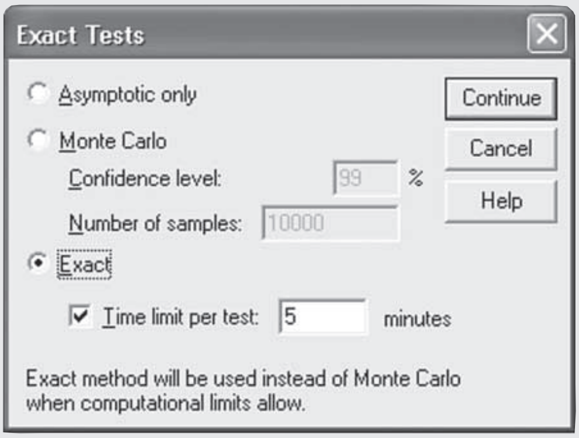
[a] Computed only for a 2 x 2 table.

[b] 1 cell (25.0%) have expected count less than five. The minimum expected count is 3.70.

From the "last line" of table VII, we observe that the validity of the Pearson's Chi-Square test is violated (**1 cell** has expected count less than five), thus in this case the p-value of 0.067 for the Fisher's Exact test should be reported (and not the significant p = 0.046 of the Pearson Chi-Square), signifying no association.

For a non 2 x 2 table, we can "ask for" Fisher's Exact test by clicking the **Exact** button (at the left corner of Template I) and the following template is obtained:

Template IV. Exaxt Tests.



○ Asymptotic only

○ Monte Carlo

Confidence level: 99 %

Number of samples: 10000

● Exact

☑ Time limit per test: 5 minutes

Exact method will be used instead of Monte Carlo when computational limits allow.

Tick the **Exact** option. The computation for this Fisher's Exact test is quite "extensive" and sometimes for a 4 x 6 table, say, most likely the Pearson's Chi-Square will not be valid as there's a high probability for some of the cells to have small frequencies. After a couple of minutes' computation, the only "answer" we get from the Fisher's Exact test is "Computer memory not enough!" What should we do?

If the p-value of the "violated" Pearson's Chi-Square test is large or very small, we have no worries as the p-value of the Fisher's Exact would not be so different. The only time we have to worry is when this "violated" Pearson's p-value is hovering around 0.04 to 0.06 (and the Fisher's Exact test did not help), then it is recommended to seek for the help of a biostatistician!

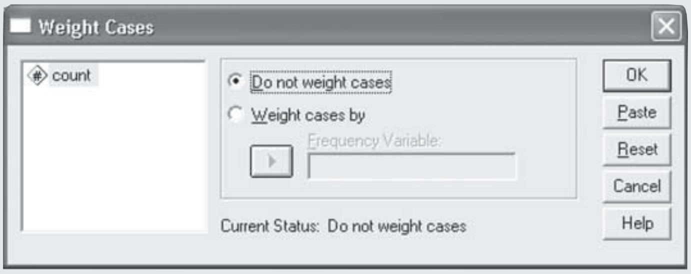
There are instances where we do not have the raw data (as given in Table I) available but only the crosstabulation Table II (perhaps appearing in a publication) and we are interested to perform the Chi-Square test. In this case, we have to set up the dataset as shown in Table VIII (refer to Table II for the corresponding frequencies).

Table VIII. SPSS data structure for a crosstabulation table.

Gender	Snoring	Count
Male	No	41
Male	At times	26
Male	Frequent	6
Male	Always	23
Female	No	58
Female	At times	31
Female	Frequent	6
Female	Always	9

Before we carry out the sequence of steps as discussed above for performing the Chi-square test, we have to "inform" SPSS that this time each row is not a subject but the total number of cases are being weighted by the Count variable. In SPSS, go to ***Data, Weight Cases*** and the following template appears:

Template V. Weight Cases.



● Do not weight cases
○ Weight cases by

Current Status: Do not weight cases

Click on the **Weight cases by** and bring the Count variable into the **Frequency Variable box**; then perform the sequence of steps for a Chi-Square test as described above.

Measuring the Strength of an Association (only for 2 x 2 tables).

The magnitude of the p-value does not indicate the strength of association between two categorical variables as we know that this value is dependent on the sample size. To express the strength of a significant association (only for 2 x 2 tables), the **odds ratio** or the **relative risk** between the outcomes of the two groups are presented. Table IX shows the crosstabulation for Exposure and Disease.

Table IX. 2 x 2 crosstabulation for Exposure and Disease.

		Disease	
		YES	No
Exposed	Yes	a	b
	No	c	d

By definition, the Odds Ratio is given by OR = (ad)/(bc): the ratio of the odds having disease given exposed and of having disease given not exposed and the Relative Risk (RR) = a(c+d)/c(a+b): the ratio of the probabilities of having disease given exposed and having disease given not exposed.

How to obtain the odds ratio and relative risk from SPSS? From template III, besides ticking on the **Chi-square** option, tick the **Risk** option too. Tables X – XI show the 2 x 2 crosstabulation and the Risk estimates for a exposure/disease example:

Table X. Crosstabulation table for Exposure and Disease example.

Exposure* Disease Crosstabulation

			Disease		
			Yes=1	No=2	Total
Exposure	yes=1	Count	30	70	100
		% within Gender	30.0%	70.0%	100.0%
	no=2	Count	10	90	100
		% within Gender	10.0%	90.0%	100.0%
Total		Count	40	160	200
		% within Exposure	20.0%	80.0%	100.0%

p<0.001 (Pearson Chi-Square)

Table XI. Risk estimates for Exposure and Disease example.

Risk Estimate

		95% Confidence Interval	
	Value	Lower	Upper
Odds Ratio for Exposure (yes/no)	3.857	1.767	8.422
For cohort Disease = yes	3.000	0.551	5.803
For cohort Disease = no	.778	.673	.898
N of Valid Cases	200		

There's a significant association between Exposure and Disease ($p<0.001$). Looking at Table XI, the Odds Ratio for an Yes/No Exposure of having Disease (the 1st column of Table X) is 3.857 (95% CI 1.767 to 8.422) which is also the OR for the No/Yes Exposure for having No Disease.

The Relative Risk is obtained from the cohort Disease = yes or no. For cohort Disease = yes, the Relative Risk between Exposure and non-exposure is 3.0 and is 0.778 for the cohort Disease = no. This interpretation of the results is rather "straightforward" because of the way we set up the crosstabulation table. Observe that the codings for "yes = 1" and "no = 2", and SPSS will display the "yes" first and then the "no". What if we have coded "yes = 1" and "no = 0" for Disease?

Table XII

Exposure* Disease Crosstabulation

			Disease		
			No=0	Yes=1	Total
Exposure	yes=1	Count	70	30	100
		% within Gender	70.0%	30.0%	100.0%
	no=2	Count	90	10	100
		% within Gender	90.0%	10.0%	100.0%
Total		Count	160	40	200
		% within Exposure	80.0%	20.0%	100.0%

Table XIII

Risk Estimate

		95% Confidence Interval	
	Value	Lower	Upper
Odds Ratio for Exposure (yes/no)	.259	.119	.566
For cohort Disease = no = 0	.778	.673	.898
For cohort Disease = yes = 1	3.000	1.551	5.803
N of Valid Cases	200		

There will be no change in the p-value of the association but from Table XIII, the OR presented now is for Yes/No Exposure of having No Disease (the 1st column of Table XII) is 0.259 (which is just the reciprocal of 3.857!).

For a non 2 x 2 table, if a significant association exists, we may want to find out where the differences are. Let's consider the example of Snoring status and Race.

Table XIV. Crosstabulation of Race and Snoring status.

RACE* Snoring status Crosstabulation

			Snoring status		
			Yes	No	Total
Race	Chinese	Count	47	64	111
		% within RACE	42.3%	57.7%	100.0%
	Indian	Count	9	3	12
		% within RACE	75.0%	25.0%	100.0%
	Malay	Count	43	27	70
		% within RACE	61.4%	38.6%	100.0%
	Others	Count	2	5	7
		% within RACE	28.6%	71.4%	100.0%
Total		Count	101	99	200
		% within RACE	50.5%	49.5%	100.0%

p = 0.013 (Fisher's Exact test).

There's an association between Race and Snoring status (p=0.013) and from Table XIV, it's not obvious where this association is. Since Race is a nominal categorical variable, we can create four dummy variables: Chinese vs non-Chinese, Malay vs non-Malays, etc. That is the new variable Chinese has only two levels: Chinese or non-Chinese and then we perform the Chi-Square test using these four dummy variables with Snoring status.

Table XV shows the crosstabulation for the Chinese and Snoring Status and the p-value for this association is 0.010 which is statistically significant even after we adjusted for the type 1 error for multiple comparison[1] ($p<0.05/4 = 0.125$). The risk estimate table XVI shows that the Chinese compared to the non-Chinese were less likely to snore (OR = 0.476).

Table XV. Crosstabulation of Chinese vs Non-Chinese with Snoring status.

Crosstab

			Snoring Status		
			Yes	No	Total
Chinese	Chinese	Count	47	64	111
		% within Chinese	42.3%	57.7%	100.0%
	Other	Count	54	35	89
		% within Chinese	60.7%	39.3%	100.0%
Total		Count	101	99	200
		% within Chinese	50.5%	49.5%	100.0%

p = 0.010 (Chi-Square test)

Table XVI. Risk estimate for Chinese vs non-Chinese and Snoring status.

Risk Estimate

		95% Confidence Interval	
	Value	Lower	Upper
Odds Ratio for Exposure (Chinese/Other)	.476	.270	.840
For cohort Snoring status = yes	.698	.531	.918
For cohort Snoring status = no	1.466	1.083	1.986
N of Valid Cases	200		

Tables XVII and XVIII indicate that the Malays compared to the non-Malays had a higher likelihood to snore but we have to be cautious about this conclusion after we have taken into consideration the adjustment of the type 1-error for multiple comparison!

Table XVII. Crosstabulation of Malay vs non-Malay and Snoring status.

Crosstab

			Snoring Status		
			Yes	No	Total
Malay	Malay	Count	43	27	70
		% within Malay	61.4%	38.6%	100.0%
	Other	Count	58	72	130
		% within Malay	44.6%	55.4%	100.0%
Total		Count	101	99	200
		% within Malay	50.5%	49.5%	100.0%

p = 0.023 (Chi-Square test)

Table XVIII. Risk estimate table for Malay vs non-Malay and Snoring status.

Risk Estimate			
		95% Confidence Interval	
	Value	Lower	Upper
Odds Ratio for Exposure (Malay/Other)	1.977	1.093	3.576
For cohort Snoring status = yes	1.377	1.055	1.798
For cohort Snoring status = no	.696	.499	.972
N of Valid Cases	200		

There were no significant association for the Indians (p = 0.080) and the Other race (p = 0.277) with Snoring status.

MCNEMAR TEST

The McNemar test is used when we have a matched case-control study. For example, we are interested to determine whether there's any association between diabetes and AMI. One study design is to match-by-age, say, a 50-year-old diabetic with another 50-year-old non diabetic and follow them up for a length of time. Four possible outcomes could be obtained. See Table XIX (which is also the SPSS data structure for a McNemar test)

Table XIX. Possible outcomes of the matched case-control study.

Diabetic Person	Non Diabetic Person	Count
Had AMI	Had AMI	9
Had AMI	No AMI	37
No AMI	Had AMI	16
No AMI	No AMI	82

To carry out a McNemar test in SPSS is exactly the same as performing a Chi-Square test, except that at Template III, we tick the **McNemar** option. Tables XX and XXI show the crosstabulation and McNemar test respectively.

Table XX.

Diabetic* Non-Diabetic Crosstabulation				
		Disease		
		AMI = No	AMI = Yes	Total
Diabetic	AMI = No	82	16	98
	AMI = Yes	37	9	46
Total		119	25	144

Table XXI. McNemar test.

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)
McNemar Test				.005[a]
N of Valid Cases	144			

[a] Binomial distribution used.

In total, we have 144 pairs of participants. There is a significant association between diabetes and AMI (p=0.005). The McNemar test compares the observations of the discordant pairs (Diabetic having AMI and Non-Diabetic not having AMI) vs (Diabetic not having AMI and Non-Diabetic having AMI) which is 37/144 (25.7%) vs 16/144 (11.1%).

CONCLUSIONS

We have covered the analysis of both quantitative[(1)] and qualitative type of data (in this article) and table XXII summarises the various techniques available.

Table XXII. Summary of Univariate Statistical techniques.

Quantitative data		Qualitative data	
Parametric test	Non-Parametric test	Independent samples	Matched samples
1 Sample T-test Paired T-test	Wilcoxon Signed Rank test	Chi-Square test/Fisher's Exact test	McNemar test
2 Sample T-test	Mann Whitney U test / Wilcoxon Rank Sum test		
ANOVA	Kruskal Wallis test		

The next article will be Biostatistics 104: Correlational analysis.

REFERENCES

1. Chan YH. Biostatistics 102: Quantitative Data: Parametric & Non-parametric tests. Singapore Medical Journal 2003; Vol 44(8):391-6.