SMARTPsych SPSS Tutorial

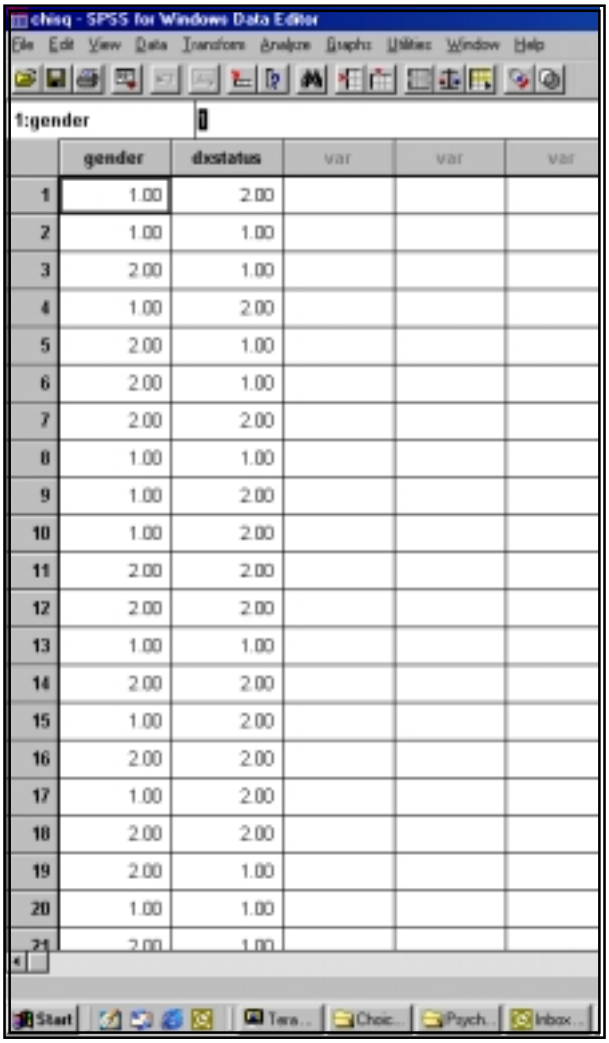
Chi-Square Test for Independence

Study Background:

You are conducting a study to see if gender is related to diagnostic status for borderline personality disorder (BPD). Due to diagnostic practices that may be gender biased, you suspect that more females than males are diagnosed with BPD. Your data are collected from outpatients at a clinic that treats suicidal behaviors. You obtain permission to review the charts of 200 patients, specifically looking at whether or not they have received a diagnosis of BPD in the past.

Are gender and BPD diagnostic status independent?

Your data look as follows:



	gender	dxstatus
1	1.00	2.00
2	1.00	1.00
3	2.00	1.00
4	1.00	2.00
5	2.00	1.00
6	2.00	1.00
7	2.00	2.00
8	1.00	1.00
9	1.00	2.00
10	1.00	2.00
11	2.00	2.00
12	2.00	2.00
13	1.00	1.00
14	2.00	2.00
15	1.00	2.00
16	2.00	2.00
17	1.00	2.00
18	2.00	2.00
19	2.00	1.00
20	1.00	1.00
21	2.00	1.00

Notice that each row represents a case (i.e., patient), and that each case is characterized by two variables (gender & dxstatus):

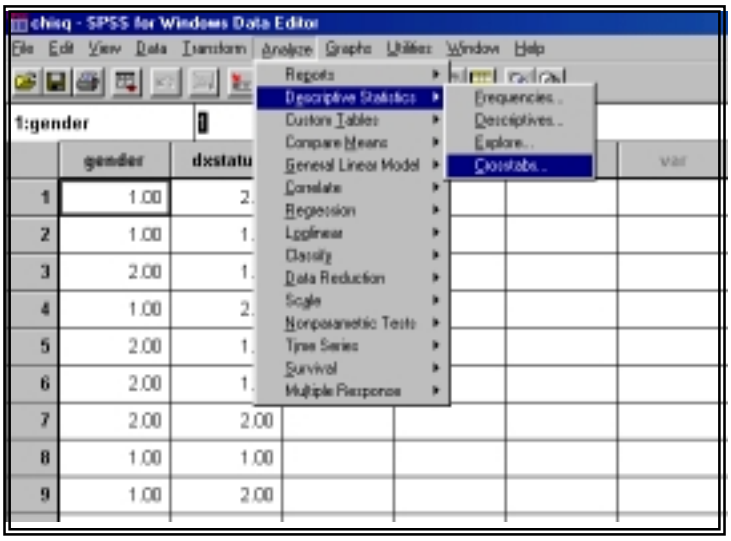
"gender" refers to the sex of the patient.

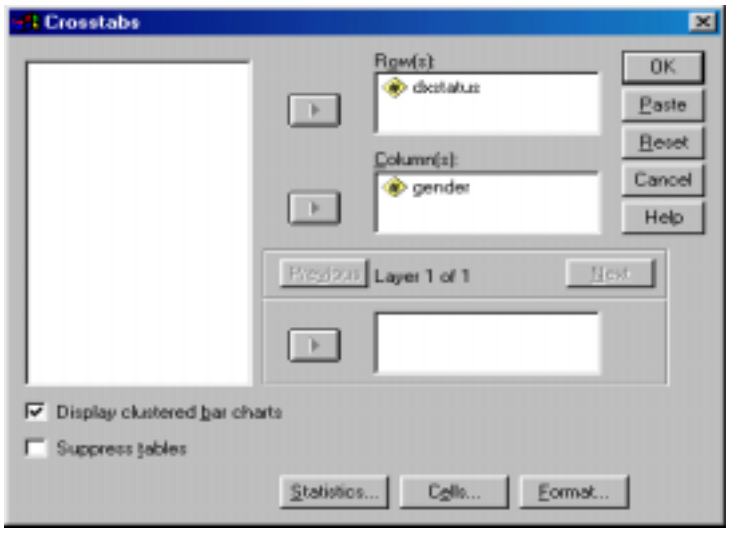
"dxstatus" refers to the diagnostic status of the patient.

To download the data file, click here. As usual, you will need to save this file to a location on your computer and open it up.

Again, you are mainly interested in whether BPD diagnostic status is independent of gender. How do you go about looking at whether two categorical variables are independent of each other? You will have to perform a Chi-Square test for independence using SPSS's Crosstabs option. You do this by selecting:

ANALYZE / DESCRIPTIVE STATISTICS / CROSSTABS

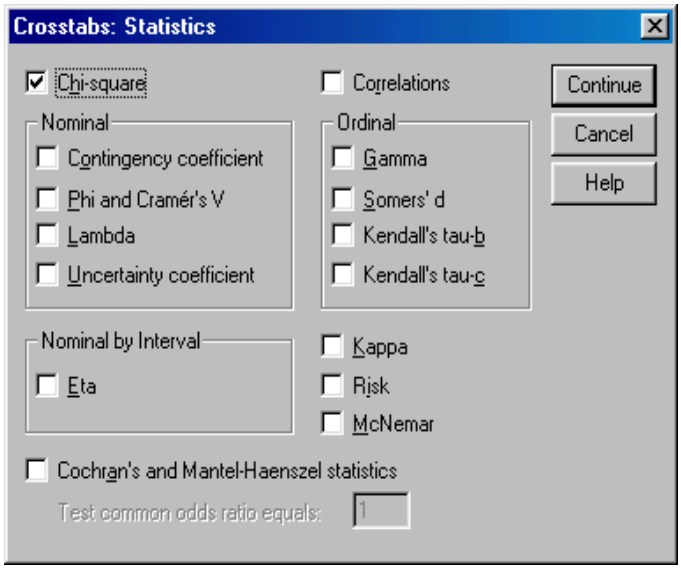




In the **Crosstabs** dialogue box, enter in one of your two variables in the Row(s) box, and the other variable in the Column(s) box. Note that it does not matter which variable goes in which box. For example, I could have entered "gender" in the Row(s) box and "dxstatus" in the column(s) box.

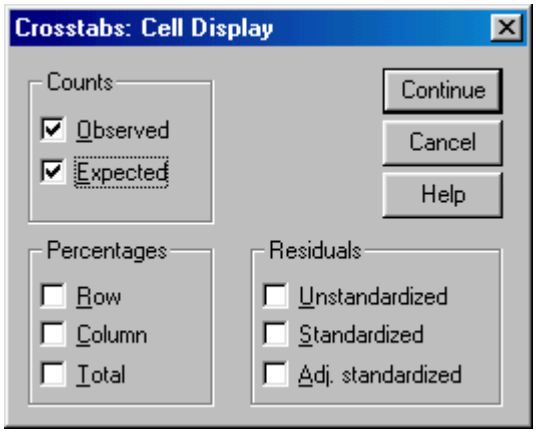
Because there are two levels of "gender" and two levels of "dxstatus", there are four possible "gender" and "dxstatus" combinations. By selecting the "display clustered bar charts" option, SPSS will produce the bar graphs of the counts (number of cases) for each of the four cells.

Next, click on "Statistics"



In the **Crosstabs: Statistics** dialogue box, select the "Chi-square statistic".

Next click on "Continue", and select the "Cells" option. You should get the following dialogue box:



In the **Crosstabs: Cell Display** dialogue box, SPSS's default option is only the Counts Observed, which will tell you the number of cases *observed* in each of the four possible cells. However, by also selecting Counts "Expected", SPSS will tell you how many cases you *expect* in each cell *if* the two variables are independent of each other.

Click on "Continue". You are now ready to click on "OK", and run your analysis.

Your SPSS output should look as follows:

Crosstabs

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
DXSTATUS * GENDER	200	100.0%	0	.0%	200	100.0%

Case Processing Summary:
This first output box is really relevant if you have missing values. It will tell you the number of valid cases (here N = 200), as well as the % of valid cases out of the total number (here % of valid cases = valid/total = 200/200 = 100%.

DXSTATUS * GENDER Crosstabulation

			GENDER		
			1.00	2.00	Total
DXSTATUS	1.00	Count	42	18	60
		Expected Count	31.8	28.2	60.0
	2.00	Count	64	76	140
		Expected Count	74.2	65.8	140.0
Total		Count	106	94	200
		Expected Count	106.0	94.0	200.0

DXSTATUS * GENDER Crosstabulation:
This second output box will tell you how many cases are in each of the four possible cells. This is your Observed Count, referred to simply as *Count*. Because we asked for the Expected Count as well (in our command), SPSS is also computing the number of cases you would expect to have in each of the four possible cells *if* the two variables were independent. This is referred to as *Expected Count*. Notice that that the Chi-Square statistic, because it is computed using the observed and expected counts, reflects the extent to which the two variables are not independent.

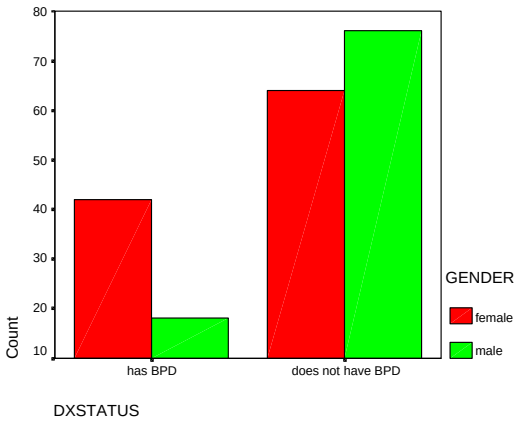
Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	9.944[b]	1	.002		
Continuity Correction[a]	8.993	1	.003		
Likelihood Ratio	10.183	1	.001		
Fisher's Exact Test				.002	.001
Linear-by-Linear Association	9.895	1	.002		
N of Valid Cases	200				

a. Computed only for a 2x2 table

b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 28.20.

Chi-Square Tests:
This third output box will give you the Chi-Square obtained, and the p-value for the test statistic. Note that SPSS gives a two-sided p-value, or significance level. This is not necessary with a Chi-Square test; divide this value by 2 to get the one-sided p-value (.002 / 2 = .001). Similar to other statistical tests we have covered in this class, a Chi-Square obtained that is greater than the Chi-Square critical value will allow you to reject the null hypothesis. Here, our null hypothesis is that the two variables are not related, that is they are independent.



Clustered Bar Charts:
This fourth output box will give you a pictorial representation of what the data look like. As you can see, there are more female patients (relative to male patients) classified with BPD. In contrast, there are more male patients (relative to female patients) classified without BPD.

What can we conclude? Is gender independent of a diagnosis of Borderline Personality Disorder?

It appears that in our sample of outpatients from a clinic specializing in suicidal behaviors (N= 200) a diagnosis of Borderline Personality Disorder is related to gender. Specifically, females are more likely to be classified with BPD than males, $\chi^2(1) = 9.94$, $p < .005$.

Note that this conclusion is consistent with our hypothesis that diagnostic practices of BPD may be gender biased. However, similar to correlational studies, these results only say that the two variables are related to each other, but it doesn't say that *why* the relationship exists. For example, it's possible that the diagnostic practices are not gender biased, rather that females are more likely than males to develop BPD.