

**60080079 Introduction to Statistical Methods**  
**Semester 2 2023-2024**  
**Handout 8**

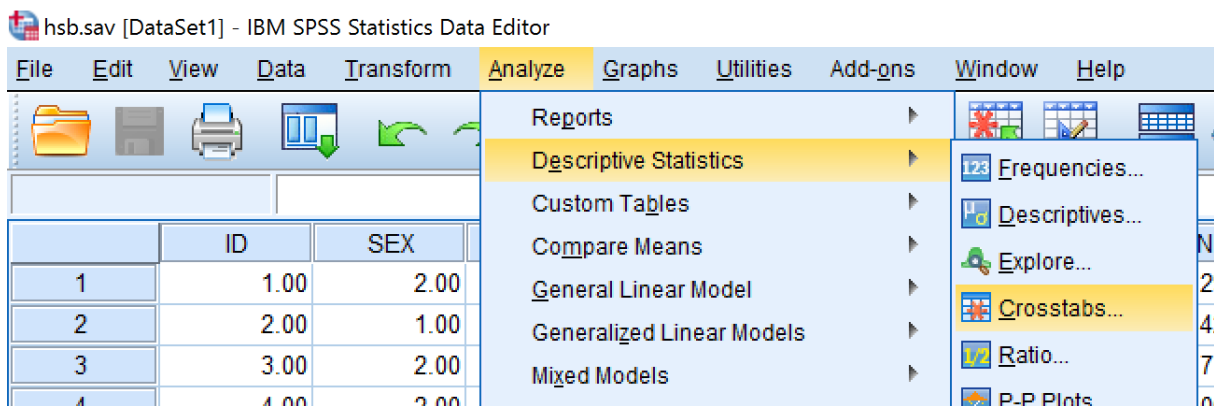
**An Introduction to Crosstabulation and Goodness-of-Fit Test in SPSS**

I. SPSS can handle crosstabulation and chi-square analysis of both raw and summarized data.

A. Raw Data

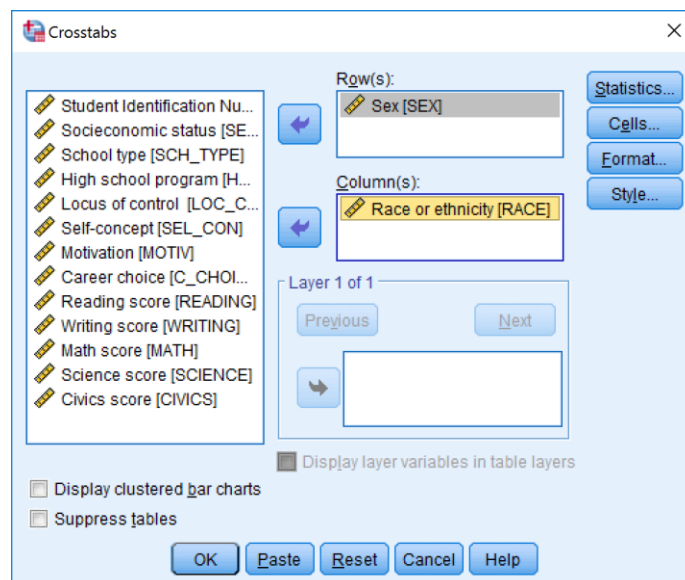
Open the **hsb.sav** file in SPSS. Crosstabulation of raw data requires two discrete variables.

**1. Analyze → Descriptive Statistics → Crosstabs**



2. In the **Crosstabs** dialog box, click in the row and column variables.

In this example, we want to examine whether or not Sex is related to Race.



3. By clicking **OK** without further options, you get a basic crosstabulation of the two variables. For the example above, we get the following table:

**Sex \* Race or ethnicity Crosstabulation**

Count

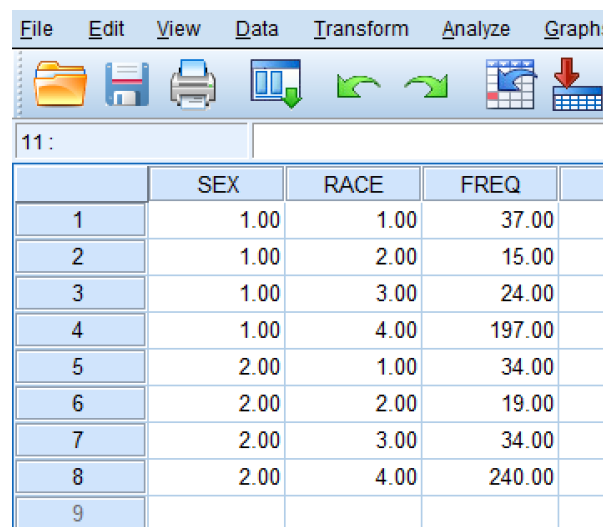
		Race or ethnicity				Total
		Hispanic	Asian	Black	White	
Sex	Male	37	15	24	197	273
	Female	34	19	34	240	327
Total		71	34	58	437	600

## B. Summarized Data

In some situations, access to the original data is not available. Instead, analysis is performed on summarized, in this case, crosstabulated data. The primary difference is how the data are set up.

0.1. Open the **hsb frequency.sav** file in SPSS. Three variables are needed to create the data in SPSS. The first two are the row and the column variables, whereas the third variable records the number of observations in each of the particular combination of row and the column variables.

In our example, we need to create a new data set with the variables SEX, RACE, and FREQ. We can use the entries from the table above to produce the following:

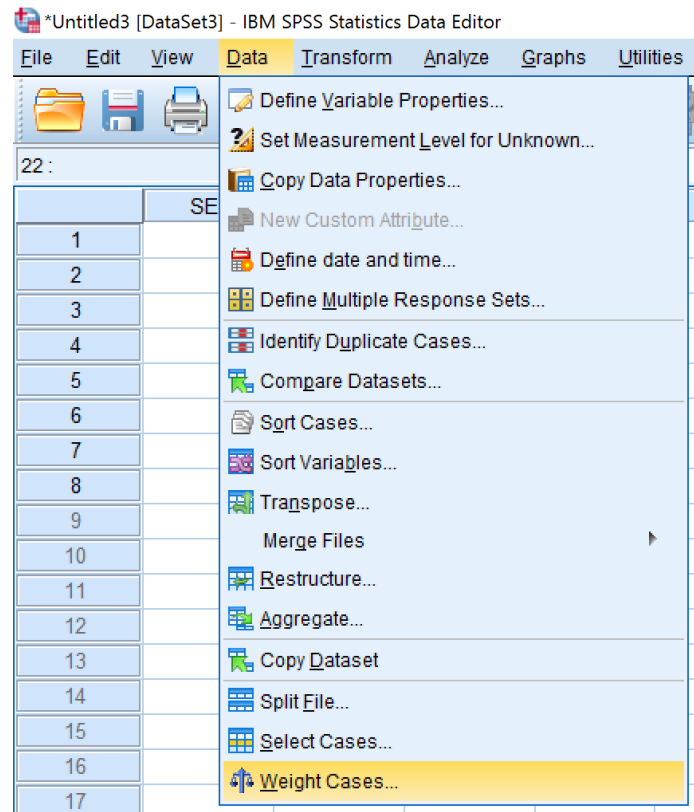


	SEX	RACE	FREQ	
1	1.00	1.00	37.00	
2	1.00	2.00	15.00	
3	1.00	3.00	24.00	
4	1.00	4.00	197.00	
5	2.00	1.00	34.00	
6	2.00	2.00	19.00	
7	2.00	3.00	34.00	
8	2.00	4.00	240.00	
9				

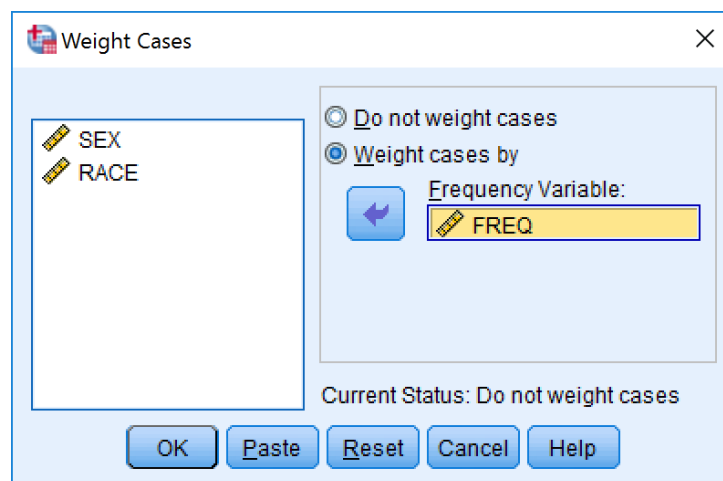
In this example, SEX has two possible values and RACE has four possible values, which result in  $2 \times 4 = 8$  combinations.

0.2 SPSS needs to be informed that the current set up is different from the previous one by identifying the weighting variable. In this example, it is FREQ.

## Data → Weight Cases



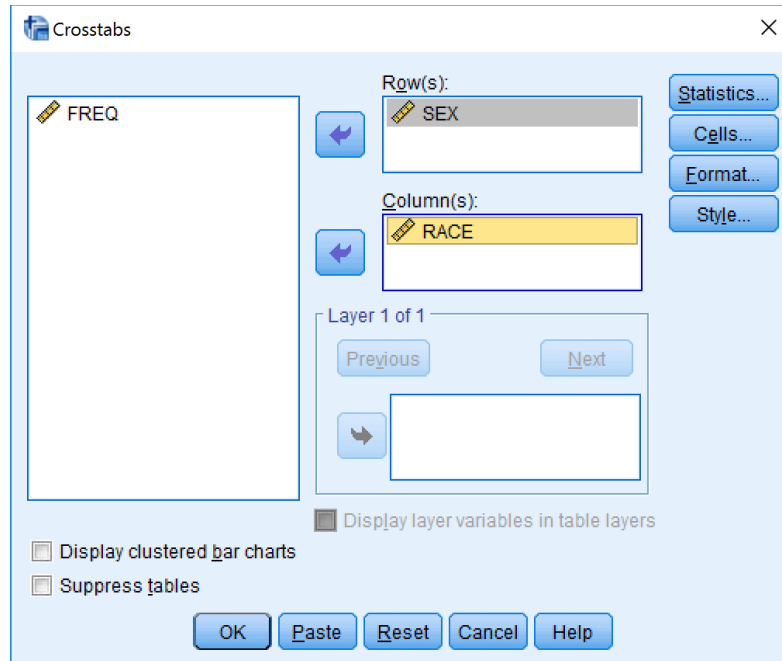
In the **Weight Cases** dialog box, choose the **Weight Cases by** option, and click in **FREQ** in the **Frequency Variable** box, then hit **OK**.



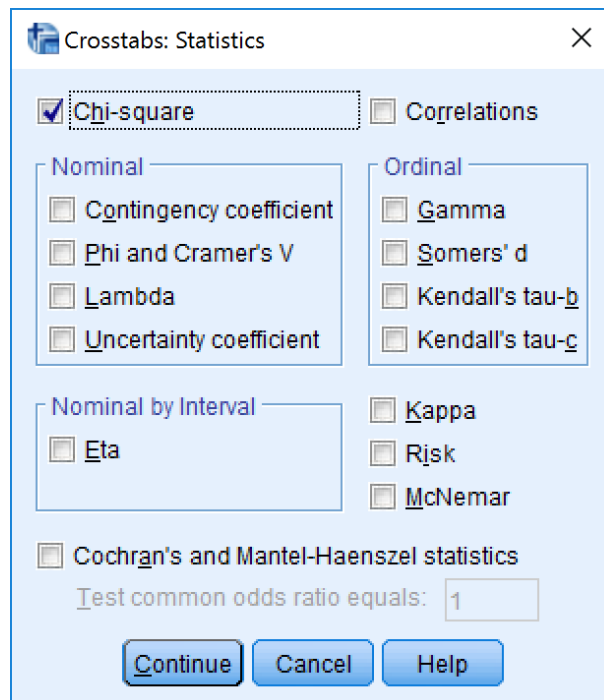
0.3. The remaining steps are similar to the steps in Part A (i.e., Steps 1-3).

### C. Relevant Options in Crosstabs

After identifying the row and column variables, but before submitting the data for analysis (i.e., between steps 2 and 3), options can be included from the **Statistics** and **Cells** buttons.



With the **Statistics** options, a **Chi-square** analysis can be requested.



With the **Cells** button, **Expected** counts can be requested in addition to the **Observed** counts (default) under **Counts**.

Different types of **Percentages** and **Residuals** (i.e., observed – expected) can also be requested.

Choosing the **Expected** and **Chi-square** options will yield the following results:

SEX * RACE Crosstabulation						
			RACE			
			1.00	2.00	3.00	4.00
SEX	1.00	Count	37	15	24	197
		Expected Count	32.3	15.5	26.4	198.8
	2.00	Count	34	19	34	240
		Expected Count	38.7	18.5	31.6	238.2
Total		Count	71	34	58	437
		Expected Count	71.0	34.0	58.0	437.0

Note: These are the expected counts under the null hypothesis that the two variables are not related.

It appears that the expected counts are close to the observed counts.

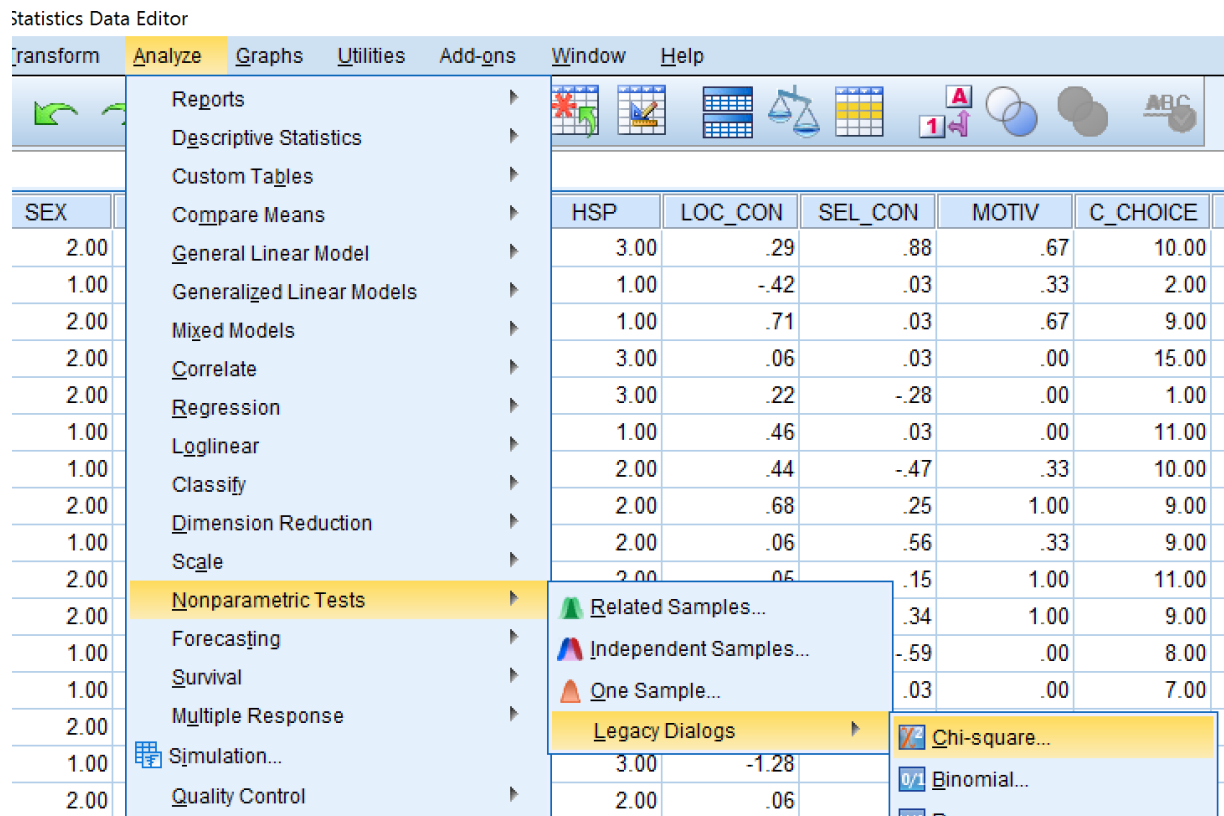
Chi-Square Tests			
	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	1.706 <sup>a</sup>	3	.636
Likelihood Ratio	1.703	3	.636
Linear-by-Linear Association	.726	1	.394
N of Valid Cases	600		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 15.47.

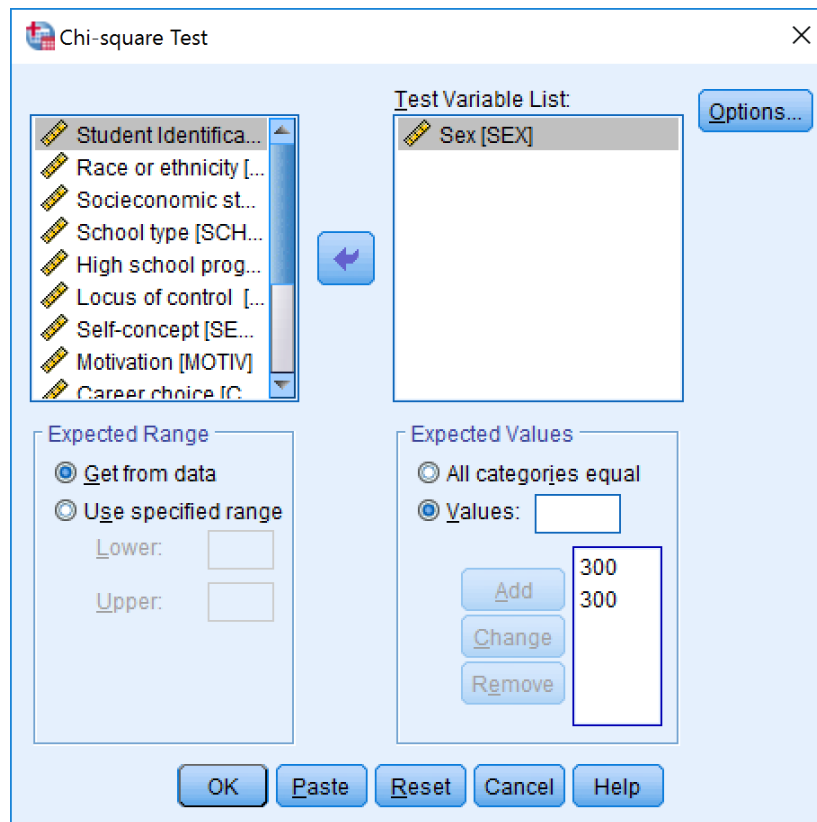
The (Pearson) Chi-square test shows a p-value of 0.636, which indicates that SEX and RACE are independent – the distribution of RACE is the same for Male and Female.

II. SPSS can also carry out goodness-of-fit test for a single categorical variable.

### Analyze → Nonparametric Test → Legacy Dialogs → Chi-square



We want to examine whether the proportions of male and female high school students are equal.



Click in SEX in the **Test Variable List** box.

We have two way of carrying out the null hypothesis the  $p_{\text{Male}} = p_{\text{Female}} = .5$  using the **Expected Values** option.

First, choose the **All categories equal** option.

Or second, specify the **Values** (i.e., expected counts) under the null hypothesis. In this example,  $n = 600$  so the expected counts are 300 for **Sex = 1** (Male) and 300 for **Sex = 2** (Female).

Note that, although this is not the case with the current example, the order by which the expected values are entered matters.

Below is the output from this analysis.

Sex			
	Observed N	Expected N	Residual
Male	273	300.0	-27.0
Female	327	300.0	27.0
Total	600		

The observed numbers of male and female students in the sample appear uneven.

**Test Statistics**

	Sex
Chi-Square	4.860 <sup>a</sup>
df	1
Asymp. Sig.	.027

a. 0 cells (0.0%) have expected frequencies less than 5. The minimum expected cell frequency is 300.0.

The p-value of .027 leads us to conclude that the proportions of male and female high school students in the population are not equal.