In the last section, we looked at commands that gave us important information about numeric variables. In our cheating dataset, we have other numeric variables that are of a slightly different nature. Take *gender* for example. We know that the variable actually contains the values 0 and 1. So it is numeric. When we go to the data editor we do not see these numeric variables because we had applied labels to them, but the fact remains that they are numeric. In order to see that run the **codebook** command for *gender*:

codebook gender

The output clearly states that the variable is numeric, and that the range of values it takes is from 0 to 1. We are also told that the unit of the variable is 1, which means that there are no decimal points. Hence, the variable either takes the value 0 or the value 1. We also see that we have 15 missing values in the 818 observations. What is different in the output however is that Stata does not display the mean, the standard deviation, or the percentiles. Stata knows that this is a categorical variable. In the case of these variables, we are not really interested in the mean, although it can be calculated. What we are really interested in is how many zeros and how many ones are in the dataset. In other words, how many males and how many females. This is why when the **codebook** command is used with this type of variable, instead of the mean and standard deviation, Stata displays a tabulation. Looking at this tabulation, we see that there are 480 occurrences of the value 0 and 323 occurrences of the value 1. Because we have applied labels to these values, Stata also tells us that the zeros correspond to males and that the ones correspond to females.

Can we use the **summarize** command? Sure, since *gender* is stored as numeric values:

summarize gender

Since there are more males than females, and since males have been assigned a value of zero, we can see that the average of the variable gender is smaller than 0.5 (closer to zero than it is to one). However, the best command to use with this type of variables is the **tabulate** command. Run the following:

tabulate gender

Stata produces a very nice table that shows us that the variable *gender* takes on two values, male and female, and the frequency of each occurrence. Stata also goes ahead and calculates the percentages for us. Now we know that 59.78% of the observations are male and that 40.22% are female. Notice that the total number of observations is 803 since there are 15 missing values. If we wanted to see the numeric values instead of the labels, we could type:

tabulate gender, nolabel

Usually we are interested in the labels, but sometimes we would like to see how the variable was coded.

The real power of the **tabulate** command however is when it is used for two categorical variables at the same time. We actually have other categorical variables in our dataset. For example, the variable *level*:

codebook level

The variable records in which year each student is and can take on four different values, ranging from 0 to 3. Try using the **tabulate** command on this variable:

tabulate level

Stata produces a table that tells us the frequency and percentage or each occurrence. Now use both *gender* and *level* in the same command:

tabulate gender level

The table that we now see is a very useful table because it tells us how many males are in each level and how many females are in each level. We now know that there are more males in all levels. Usually it is better to look at percentages because they make comparisons easier. We can tell Stata to give us the percentages in addition to the frequency:

tabulate gender level, column

Here we have used the **column** option. This option tells Stata to calculate the percentages in each column. As you can see the sum of the percentages in each column must be 100. Looking at the output, we see that 40% of freshman students are females and 60% are males. We can also tell Stata to calculate the row percentages:

tabulate gender level, row

Here we see that the sum of the percentages in each row add to 100. This output tells us that 3.77% of males are in the freshman year and that 41.42% of males are in the senior year. So basically, each output gives us different information, and it is the researchers job to determine what information he or she is looking for.