Visualizing data in Stata – Section 3.1
Najib Mozahem

The previous section covered how we can look at individual variables. Usually in data analytics, we are interested in understanding differences between different groups of people. While it is useful to know the distribution of the variable *gpa*, what might be more illuminating is investigating whether there are differences in the variable across genders. Do males have a higher GPA? If this was the case, then we would expect the average of the variable for the males to be shifted more to the right. Perhaps the average of both genders are the same but males display more variability. This would mean that the tails of the distribution for males would be longer. We can also investigate difference between students who are studying engineering and those who are studying business.

What about the variables *think*, *engage*, and *other*? In the exercises included in the previous section, you were asked to inspect the distribution of these variables. What if we compared the distribution of those who had taken a business ethics course and those who hadn't, with regards to the variable *engage*? Remember that higher values of *engage* indicate that the student reports a higher level of engaging in academic misconduct. We would expect that the distribution of this variable for students who had taken the course business ethics would be to the left of the distribution of students who had not yet taken the course. This would mean that taking this course will result in students engaging in such behavior less.

Studying group differences is one of the main reasons why statistics and data analytics are used in the social sciences. This section will deal with investigating group differences when observations are divided using a single variable. Later, in another section, we will see how we can further divide the observations using more and more variables.