Visualizing data in Stata – Section 1.2
Najib Mozahem

Attached to this lecture, you will find a copy of the dataset that will be used throughout this course. The name of the file is *thedata.dta*. Download the file and save it in whatever folder that you want. After that, load the dataset into Stata:

use thedata

In my case, I have stored the dataset in my current working directory, so I can load the dataset by just referring to its name. Depending on where you stored the file, you might have to use the full path name in order to tell Stata where to locate the file.

Once you load the data you will notice that the label of the dataset is "Student perceptions about academic misconduct". We conducted a survey in several universities in which we asked students some questions in order to better understand their views and behavior towards academic misconduct. Some of the data in this dataset is true, i.e. I did not tamper with it. Other variables however have been created just for the sake of illustration. This means that you should not give too much meaning to all of the results that we obtain when we look at the dataset.

Execute the following in order to get more information about the dataset:

describe

We see that there are 819 observations in the dataset. Looking at the variable list we see that we have a variable that tells us whether the student was enrolled in the business or the engineering school (we only surveyed students from these two schools). We also see that the variable *major* records in which particular major the student was enrolled. We also have information about the number of credits completed so far, gender, and the cumulative GPA so far. The variable *attendance* records the average grade that the student got in his or her last semester on participation and attendance. In all courses instructors assign grades to attendance in order to motivate students to come to class. There is also a variable that stores the students grade on the English course. The survey was conducted in a non-English speaking country, but all universities included in it where English teaching institutes. Students studying business and engineering are required to take a standard English course. The reason why this variable is there is that we wanted to see whether students who did not have a good grasp of the English language suffered in their major courses due to communication problems. The dataset also has a variable that measures how many courses has a student withdrawn from. We also have a variable that measures the yearly income in U.S. dollars of the student's household. This is followed by the three variables *ethical*, *fair*, and *grades*. These variables record the students' responses to the questions "Cheating is ethically wrong", "Cheating is not fair", and "Grades are more important than knowledge". We then have the three variables *liecustomer*, *liecoworker*, and *illegal*. These variables are the students' responses to the questions "Sometimes it is necessary to lie to a customer to protect the company", "Sometimes it is necessary to lie to a co-worker to protect the company", and "Sometimes it is necessary to do something that is illegal". The variable *course* records whether the student has already taken the course "Business ethics". In all universities included in the survey, all business students must take this course at one point, while engineering students can take it as an elective. The variable *university* records which university the student attends.

Finally, the last three variables measure three different dimensions with regards to academic misconduct. The variable *think*, measures to what extent the students believes that certain acts are considered as academic misconduct. For example, a student might think that working on an online quiz with his or her friend is not considered as academic misconduct. Other students might disagree. The higher the value of the variable *think*, the more likely that the student is to classify a certain situation as cheating. The variable *engage* on the other hand measures to what extent the student has engaged in certain acts that might be classified as academic misconduct. Again, the higher the value, the higher the level of engagement in such acts. Finally, the variable *other* measures the student's belief with regards to the extent that other students are engaging in acts that can be classified as academic misconduct. For our sake, it is not important to understand how we got these variables. These variables were calculated from the responses of a series of 24-questions. Our dataset does not include these questions, it just includes the students' scores on these three variables. What we need to know is that the higher the value of *think*, *engage*, and *other*, the more the student thinks that certain acts are cheating, the more the student himself or herself has previously engaged in the act, and the more that he or she believes that other students are engage in the act.

The dataset is already formatted in a way to make reading it easier. For example, run the following command:

tabulate ethical

You will see that there were three possible responses: disagree, neutral, and agree. If you run the same command on the other categorical variables that have a value label assigned to them, you will notice that *college* has two values (business or engineering), *gender* has two values (male or female), and that *major* has eight values (Management, Human Resources, Marketing, Banking and Accounting, Electrical and Computer, Civil, Mechanical, and Biomedical). You will also notice that with regards to the variable *course*, students have either taken the course or they have not yet taken it. Students who were currently taking the course were asked to state that they have not yet taken it because the surveys were distributed at the start of the semester.

So this is our dataset. Our job in this course is to extract information from it using the graphical tools that are available in Stata.