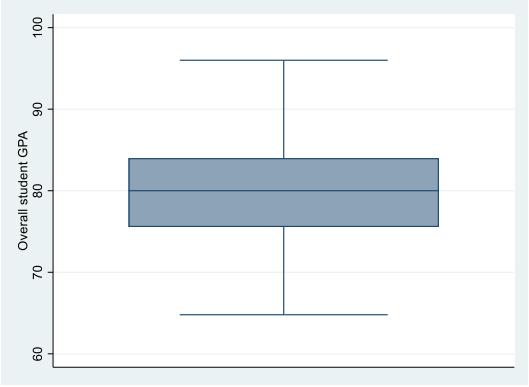
Visualizing data in Stata – Section 2.4
Najib Mozahem

In the previous lecture, we saw how we can use quantile plots to study the distribution of a variable. Quantile plots calculated all the quantiles of a variable and plotted these values. However, sometimes we are not interested in all quantiles. Sometimes all a researcher wants to know about a variable are the values of some of the quantiles. This information is summarized in what is referred to as Box plots. This plot uses only five statistics of the variable: the lower limit, the 0.25 quantile, the median (the 0.5 quantile), the 0.75 quantile, and the upper limit. This is best understood using an example. Again, we use the variable *gpa*:

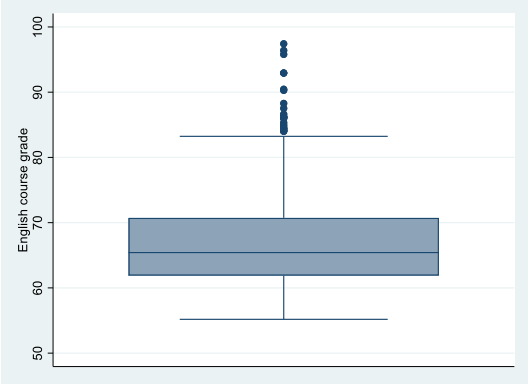
graph box gpa



The blue box represents the area where 50% of the observations lie. The top of the box is the 0.75 quantile, the bottom of the box is the 0.25 quantile, and the horizontal line in the middle of the box represents the median. The two horizontal lines that extend beyond the box represent the upper and lower limits. These are not the minimum and maximum values. The upper and lower limit help us determine whether there are outliers. Outliers are points that are very distant from the rest of the data. If there were outliers, they would be displayed as circles. Looking at our graph, we see that there are no outliers. The graph tells us that the median for *gpa* is around 80%, that the 0.25 quantile is around 75%, and that the 0.75 quantile is around 85%. The graph also tells that the distance between the 0.75 quantile and the median is equal to the distance between the median and the 0.25 quantile. These results are not surprising given that we have already seen that this variable, to a large extent, follows a normal distribution.

Let us now produce the box plot for the variable english:

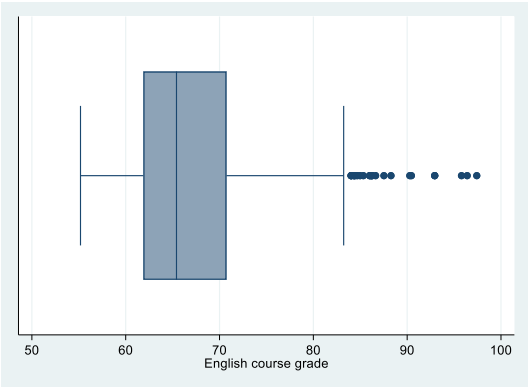
graph box english



In this case we see that we have outliers, and that all of the outliers are a result of observations having a larger than expected value. We also see that the distance between the 0.75 quantile and the median is considerably larger than the distance between the median and the 0.25 quantile. These results are not surprising since we have already seen that the variable *english* is skewed to the right. This means that a large number of observations are clustered on the left and then we have some observations extending to the far right. This results in the 0.75 quantile being further away, and it also results in outliers on the right of the distribution.

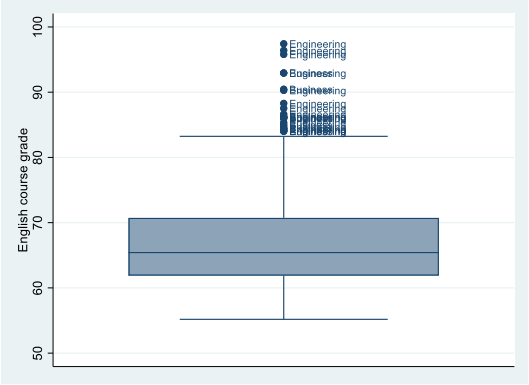
We can also ask Stata to plot a horizontal box plot:

graph hbox english



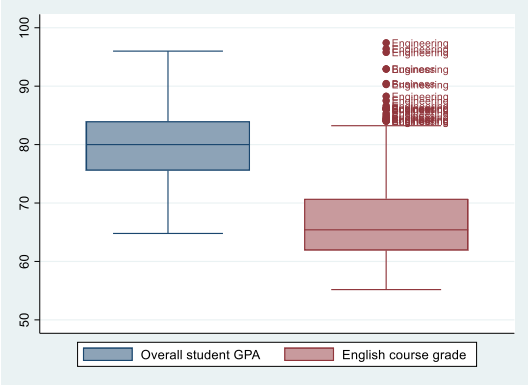
One useful option to use when plotting a box plot is to ask Stata to identify the outliers by labelling them. Perhaps we want to know in which college are these outliers? This is accomplished using the following command:

graph box english, marker(1, mlabel(college))



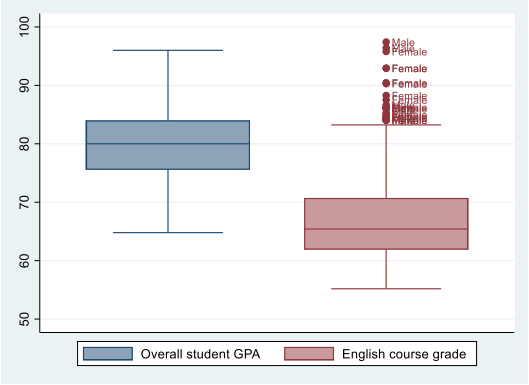
We have used the **marker()** option to tell Stata that we want to do something to the markers on the graphs, i.e. the points that represent the observations. The number "1" is used to tell Stata that we are referring to the observations of the first variable. In this case we only have one variable, but it is possible to include two variables in the command:

graph box gpa english, marker(2, mlabel(college))



In this command we asked Stata to plot two box plots, one for the variable *gpa* and the other for the variable *english*. Since english was listed second, we used the number "2" to tell Stata that we are referring to the second variable. Inside the **marker()** option, we used the **mlabel()** option to label the observations. We specified the *college* variable to tell Stata to label the observations by the value of *college*. We see that the top three outliers are all from the college of engineering. We can easily ask Stata to label the observations by another variable:

graph box gpa english, marker(2, mlabel(gender))



In this case, we told Stata to label the observations by their genders.

Sometimes, instead of labelling the outliers, you might want to disregard them completely. This is accomplished by specifying the nooutsides option:

graph box english, nooutsides