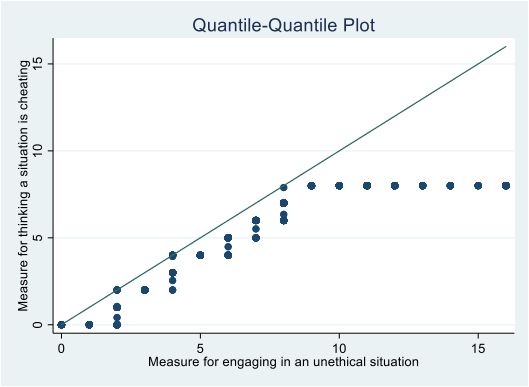
Visualizing data in Stata – Section 6.2
Najib Mozahem

Previously in the course, we used quantile plots to study the distribution of a certain variable. We also used quantile-quantile plots to compare the distribution of a variable across two groups. In this section, we will use quantile-quantile plots to compare the distributions of two variables.

It would be interesting to compare the distributions of the variables *think* and *engage*. The reason is that if *engage* took on higher values than *think*, then what this means is that students are more willing to engage in cheating than they are willing to label it as cheating:

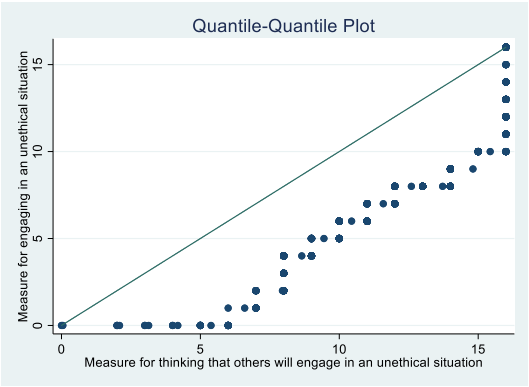
qqplot think engage



The reason why there are consecutive horizontal dots near the end of the graph is that the maximum value of *think* is 8 while the maximum value of *engage* is 16. So once we reach the quantile that includes the maximum value of *engage*, the value of the quantile no longer changes, since the maximum has been reached. Prior to that point however, we notice that the dots tend to be below the diagonal line. Given that the variable *think* is plotted on the y-axis, this means that the quantiles for *engage* are larger. What this indicates is that students report a higher level of engaging in certain acts compared to their tendency to label these acts as cheating. This makes sense. A student wouldn't feel comfortable in saying that an act is cheating if he or she engages in it.

Let us now compare the distributions of the variables *engage* and *other*:

qqplot engage other

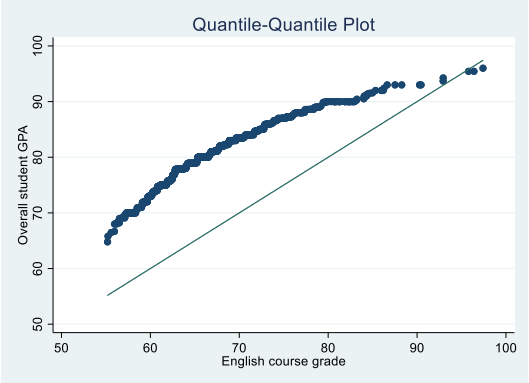


All the dots lie below the diagonal line and the distance is actually large. What does this mean? The variable *engage* is on the y-axis while the variable *other* is on the x-axis. Therefore, the plot

is telling us that the quantiles for *other* are higher (much higher) than the quantiles of *engage*. This is a very important result, because it clearly shows that an individual student believes that others cheat much more than he or she does. This is a famous mechanism for justifying ones action. You believe that everyone else is doing it, and that they are doing it much more than you, so you go ahead and partake in the act.

Finally, we will illustrate the quantile-quantile plots by comparing the distribution of the variables *gpa* and *english*. This would help us understand if the students have higher grades on English in comparison to their overall GPA, or do the grades obtained on English push down their GPA.

qqplot gpa english



The graph indicates that the values of *gpa* tend to be higher than the values of *english*. This result is not surprising given that the data was collected in countries where English is not the first language. It has been my experience that students always complain that their grades on the English course lowers their overall GPA.

As you can see, quantile-quantile plots are a very valuable tool in comparing the distributions of variables. Quantile plots are very helpful in general and it is unfortunate that they are not used more often. So far in this course we have used them to study the distributions of single variables, to compare the values of a single variable across two groups, and finally to compare the distributions of two variables.