Visualizing data in Stata – Section 3.3
Najib Mozahem

In the previous lecture we compared the histograms of two groups in order to investigate whether there were differences in the distributions of a variable between these two groups. This comparison is useful but it is difficult to interpret, even when we create the kernel density graph. The reason why it is difficult to interpret is that it is left to us to look at both graphs side and side and to determine where do these graphs differ and where do they seem the same. A more objective and direct way to compare the distributions of these two groups is to use quantile-quantile plots. If you recall, in section 2 of this course we used quantile plots in order to see if the distribution of a variable is symmetric. We also used the **qnorm** command, which allowed us to compare the distribution of a variable with the normal distribution. In this lecture, we will use quantile plots in order to directly compare the distribution of a variable between two groups. These plots are referred to as quantile-quantile plots because they compare the quantiles of one variable with the quantile of another variable. In this lecture we will actually be comparing the distributions of the same variable but across two groups. Later on in the course, we will use quantile-quantile plots to compare the distributions of two completely different variables.

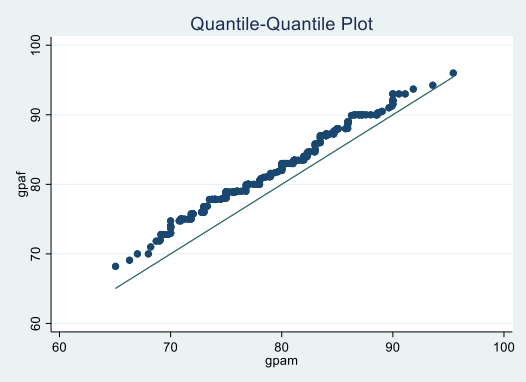
In the last lecture, we compared the histograms of the variable *gpa* for females and for males. We noted that the histogram for females appeared to be slightly shifted to the right, which indicated that females had a higher median. Let us now compare the quantile plots for *gpa* for the two genders. The command that produces the plot is **qqplot**. Unfortunately, the command does not have a **by()** option. The command only works if we specify two variables. As I mentioned, we are actually comparing the distribution of one variable, which is *gpa*. This is why in order to use this command to compare the distribution of the variable for males and females, we will need to create two new variables where the first stores the GPAs of males and the second stores the GPAs of females. This is accomplished by executing the following two command:

generate gpam = gpa if gender == 0

generate gpaf = gpa if gender==1

In the first command we create a new variable that is equal to the value of *gpa* if and only if the observation belong to a male (0 in our dataset indicates a male). The second command creates another new variable that stores the value of *gpa* if and only if the observation belongs to a female. Therefore, we now have two variables, where one is the GPA of male students and the other is the GPA of female students. We can finally use the **qqplot** command:

qqplot gpaf gpam



If you recall, when we used the quantiles to compare the distribution to a normal distribution, it was mentioned that if the distribution resembled a normal distribution then the points will lie along the line. The same logic is applied here, but instead of comparing the distribution of a variable with the normal distribution, we are comparing the distribution of two variables. Therefore, if both variables have the same distribution, the points will lie along the line. We see that this is not the case in the figure above. We actually see that the dots lie a certain distance above the line. What does this mean? Since the y-axis represents the GPA of female students, this simply means that that the quantile values of females are larger than those of males. So for example, if the GPA 70% is the 0.3 quantile of males, a GPA of 75% would be the 0.3 quantile of females. So while 30% of males have a GPA that is lower than 70, 30% of females have a GPA that is lower than 75. This means that the GPAs of females tend to be higher. Given that every point on the graph is above the diagonal line, it means that each quantiles for females is larger than the corresponding quantile for females. Therefore, the quantile-quantile plot proves what we have suspected, and that is that the distribution of GPA for males and females is not the same, with the values for females being higher.

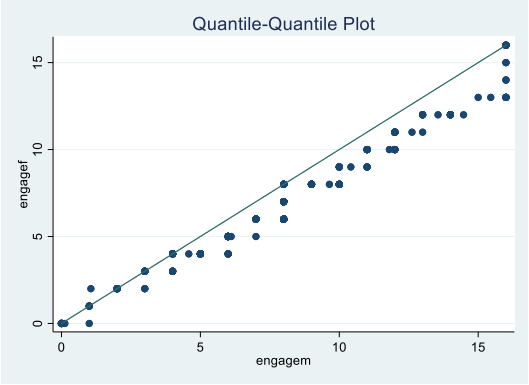
Also in the previous section we compared the value of the variable *engage* across both genders. Let us now produce the quantile-quantile plot to be able to form a better opinion. Again, we need to create the variables:

generate engagem = engage if gender==0

generate engagef = engage if gender==1

Now we create the plot:

qqplot engagef engagem



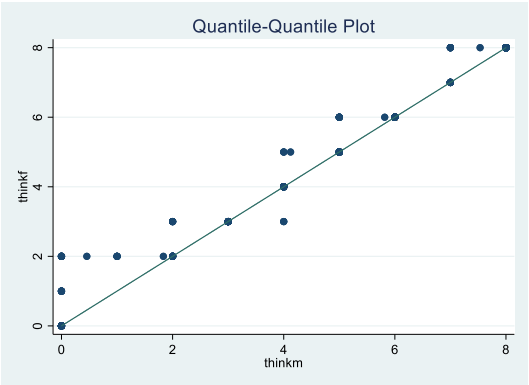
We now see that the dots lie below the line. We use the exact same logic that we used in the case of the variable *gpa*. Since the values for females are plotted on the y-axis, this indicates that the quantiles for females are less than the quantile for males. Had they been equal, the points would lie on the diagonal line. Had they been greater, then the points will lie above the diagonal line. Since they lie below the line, the quantiles for females are smaller. This means that the value of the variable *engage* is larger for males than it is for females. Hence, it seems that males report that they engage in cheating more than females do.

Again, following what we did in the last section, we will produce the quantile-quantile plot for the variable *think*:

generate thinkm = think if gender==0

generate thinkf = think if gender==1

qqplot thinkf thinkm



The plot shows that some points lie on the diagonal line, and one even lies below the line. However, the majority of the points lie above the line. Given that the value of *think* for females is plotted on the y-axis, we conclude that the quantiles for females are larger than the quantile for males. This means that females tend to have higher values of the variable.

As you can see, the quantile-quantile plot gives us more concrete evidence, instead of us having to rely on our visual ability when comparing. By plotting the quantiles of the variable across two groups, we can reach a conclusion more objectively. As I said, later on in the course, we will be

using this same plot to compare completely different variables, instead of comparing the same variable across two groups.