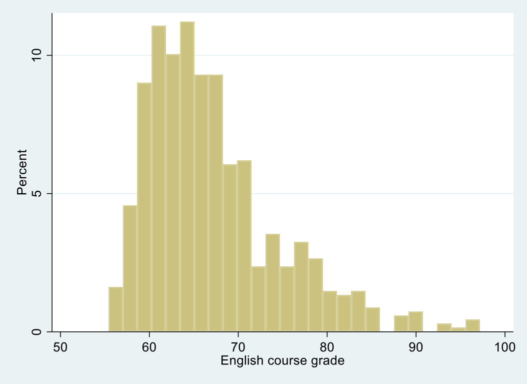
Visualizing data in Stata – Section 2.3
Najib Mozahem

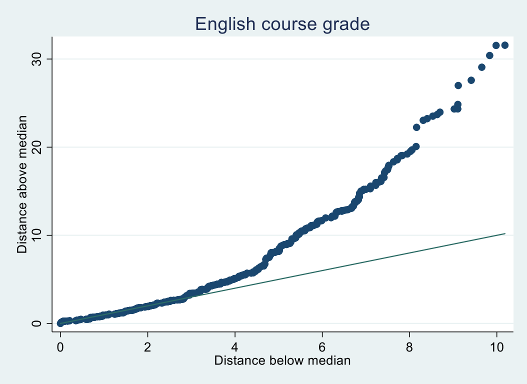
In the previous section we used histograms to look at the overall distribution of a variable. We also discussed how we can overlay the histogram with a normal distribution. In a normal distribution, the number of observations to the right of the median or middle is equal to the number of observations to the left. In addition, the variation of the observations around the mean is also the same. However, we saw that not all variables have this type of distribution. If you solved the exercises of the previous section, you should have produced the histogram of the variable *english*:

histogram english, percent



In this case, we see that the tail of the right hand side trails of more than it would if the observations were equally divided between right and left. In such a case we say that the graph is positively skewed. While histograms give us an indication of whether there is skewness or not in the data, there are other graphs that do a better job. One of these graphs is the symmetry plot. This plot helps us investigate whether the variable is symmetric or not:

symplot english

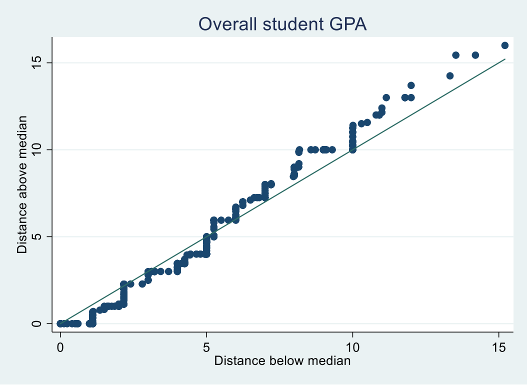


The green line represents the path of the points if the distribution was symmetrical. The idea is that if we arrange the data points in order, the data points on different sides of the median will be equidistant from the median. So the difference between the highest grade on English and the median will be equal to the difference between the lowest grade on English and the median, and the difference between the second highest grade on English and the median will be equal to the difference between the second lowest grade on English and the median, and so on. In such a case, since corresponding points are equidistant from the median, when we plot these distances, they should fall along the line y = x. Looking at our plot we see that this is not the case. Instead, what we see is that as we move forward, the distance between the points above the median and the median becomes larger than the distance between the points below the median and the median. This shows that the right tail of the graph is longer than it should be, since the points on the right are further away than the points on the left.

Why is this information useful? Simply because now we know that that highest grades in English are much higher than the lowest grades are low.

Now consider the variable *gpa*. As we saw in the previous section, when we compared the distribution of the variable to the normal curve we saw that they were pretty close. This means that the distribution of *gpa* is symmetric, since the normal curve is symmetric itself. Let us now see whether this was true or not:

symplot gpa

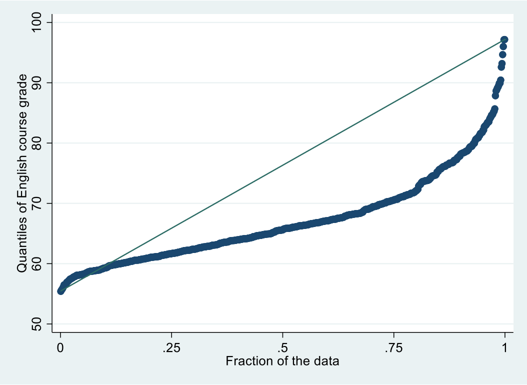


Looking at the symmetry plot, we see that the variable *gpa* is actually pretty symmetric. So now we know that in terms of grades at least, the top students do as well as the bottom students do bad.

The symmetry plot is useful when we want to know whether a variable is symmetric or not. Another plot that tells us whether the variable is symmetric is the quantile plot. Quantile plots are more important than symmetry plots because not only do they tell us whether a variable is symmetric, but they also allows us to compare two variables, something that we will do in section 3 of this course. For now, we will use the quantile plot to see whether the variable *english* is symmetrical.

First, however, you need to understand what a quantile is. Quantiles divide the data into several parts. They represent the values below which a certain fraction of the data lies. The 0.1 quantile is the value that 10% of the observations are less than or equal to. The 0.3 quantile is the value below which 30% of the observations lie. The 0.5 quantile is the value under which half of the observations lie, so it is the median. Quantile plots represents this information graphically. They tell us what percent of the observations lie below a certain value. To produce the quantile plot of the variable *english*:

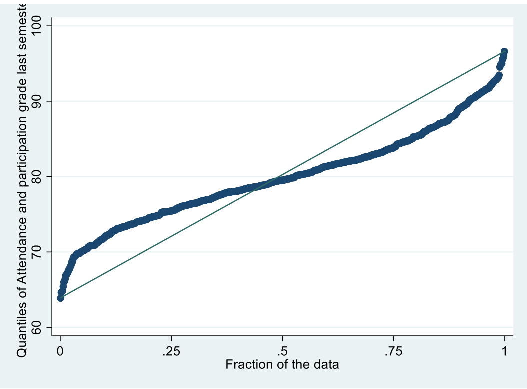
quantile english



Unlike the symmetry plot, we see that the plot now lies below the reference line. This makes sense when you think of what the quantile plot is actually plotting. The y-axis shows the quantiles and the x-axis shows the fraction of the data that lies below each quantile. If a variable is skewed to the right, as we already know that this variable is, most of the observations will lie to the left of the median with some of them trailing to the right. If the bulk of the observations lie on the left, then lower values of the variable will result in a large percentage of the observations being less than or equal to that value. So we reach the 25% quantile for example faster than we would have if the distribution is symmetric. This is why low values on the y-axis correspond to large values on the x-axis.

To further illustrate this, let us look at the quantile plot of *attendance*, since we have not looked at its symmetry plot:

quantile attendance

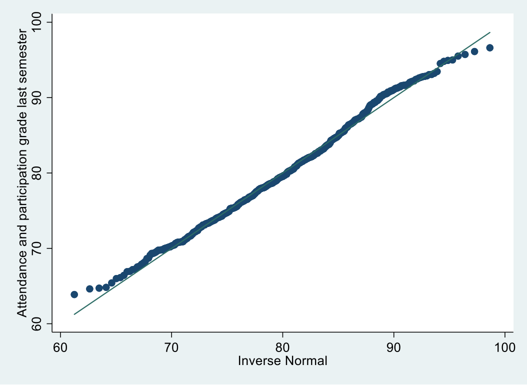


Looking at the graph, we see that the data seems to be symmetrical up to an extent. On the left we see that the graph seems to be skewed to the left, since it is taking larger values of the variable *attendance* to reach certain fractions of the data. On the right, the variable is skewed to the right since it is taking smaller values of the variable *attendance* to reach certain fractions of the data.

Unlike symmetry plots, quantile plots can be used to compare two distributions. This is because quantile plots use fractions, not absolute values. When we use fractions we can compare things. Imagine that there are two restaurants, restaurant A and restaurant B. Restaurant A has 50 employees and restaurant B has 300 employees. Now imagine that out of the 50 employees in restaurant A, 10 have a university degree, and that out of the 300 employees in restaurant B 30 have a university degree. If I say that restaurant B has more employees who have a university degree than restaurant A, I will be telling the truth, but the comparison is not fair. I can however say that 20% of the employees in restaurant A have a university degree while only 10% of the employees in restaurant B have a university degree. In this case, I reach the opposite, and more factual, conclusion, which is that restaurant A has a work force that is more educated.

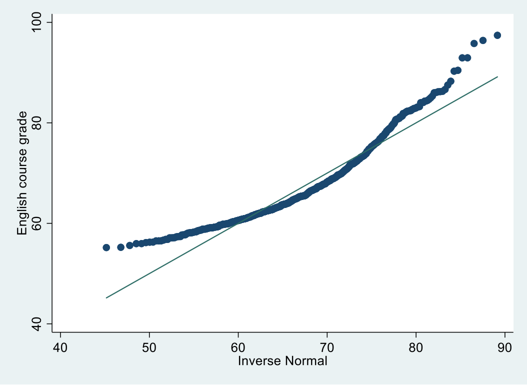
This is the same logic when it comes to why we can use quantile plots to compare two variables but not symmetry plots. This section will not cover the comparison of two variables, but it will cover the comparison of our variable with the normal distribution. As mentioned before, in happens that in many cases, we might want to know whether our variable is normal or not. In the previous lecture, we plotted the normal distribution on top of the histogram of our variable. A better way would be to compare the quantile plot of our variable with that of a normal variable. This is accomplished using quantile-normal plots:

qnorm attendance



We see that the plot lies on the reference line which indicates that the variable's distribution is actually very close to a normal distribution. We know that the variable *english* turned out to be skewed:

qnorm english



Again, we see evidence that the distribution of the variable is not similar to a normal distribution.