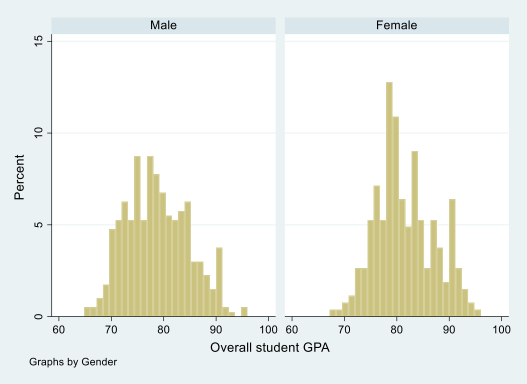
Visualizing data in Stata – Section 3.2
Najib Mozahem

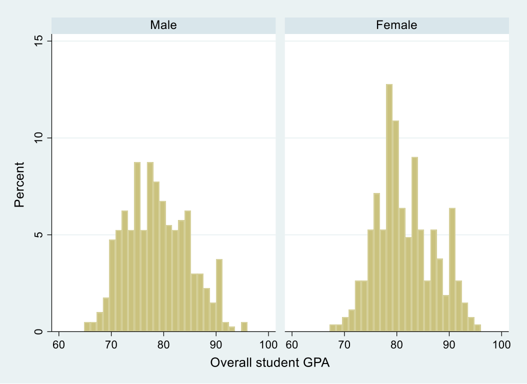
First we start with the case of histograms. To tell Stata to produce histograms for different groups of people, we need to use the **by()** option:

histogram gpa, percent by(gender)



The above command told Stata that we want a separate histogram for each group included in the variable *gender*. Since in our dataset there are two such groups, males or females, Stata produces two histograms. There is also a note at the bottom telling us that the graphs are divided by gender. If you want to eliminate this note, you can do the following:

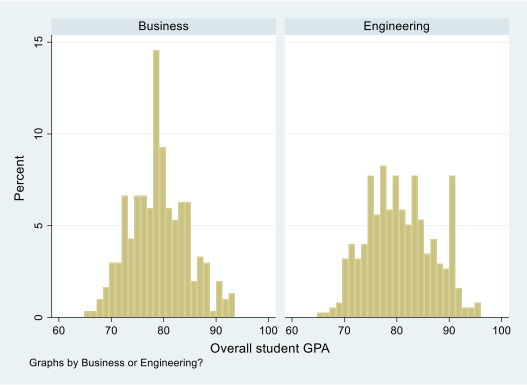
histogram gpa, percent by(gender, note(""))



As you can imagine, you customize the note by inserting whatever text you want inside the quotations. Looking at the two histograms, we see that the distribution for females is actually more to the right than that of males, indicating the females have a higher GPA. We also see that a larger percent of female students score around the median in comparison to male students.

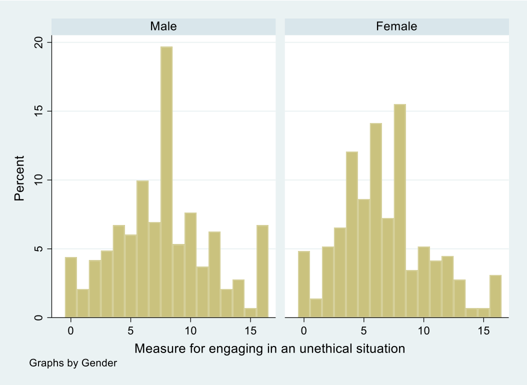
We can also divide the results based on the variable college:

histogram gpa, percent by(college)



What about the variable *engage*? It would be interesting to compare males and females:

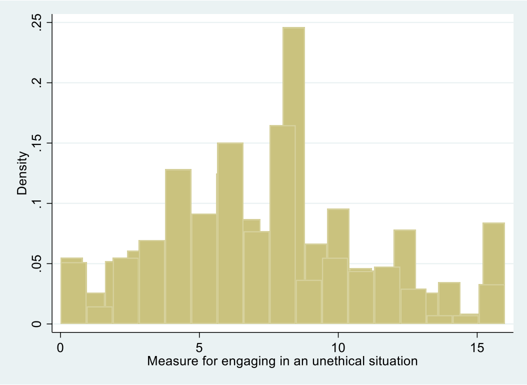
histogram engage , percent discrete by(gender)



Notice that I used the **discrete** option. This is because the variable engage only takes on integer values. You can verify this if you execute the codebook command on it. As was mentioned in a previous lecture, when the variable takes only integer values, using the **discrete** option will result in Stata calculating the width of the bars in a more intuitive way and this will result in better graphs.

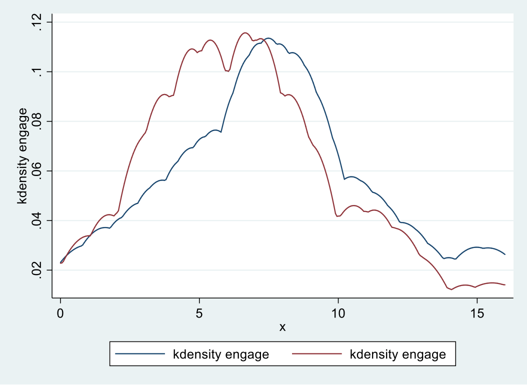
Looking at the figure, it is actually quite difficult to compare both of them. However, it would be easier to compare them if both plots were on the graph instead of being side-by-side. Using the by() option produces two separate graphs. Fortunately, there is a way to plot both on the same graph, but it requires using the **twoway** command:

twoway (histogram engage if gender ==0) (histogram engage if gender == 1)



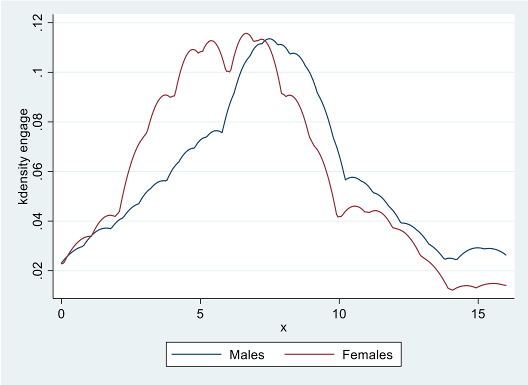
The **twoway** command is very powerful and it is the most used command when it comes to data visualization. We hadn't used it so far because there was no need. However, to create more complex graphs, we cannot keep ignoring it. Using this command we can plot as many graphs on the same axis. This makes comparison easier. We tell Stata what are the individual plots by including each in its own parentheses. In the same parenthesis, we are telling Stata to plot the histogram of the variable *engage* but only to include the observations where the value of gender is 0. In our dataset, a value of 0 for *gender* means that the individual is a male. The second plot included in the command tells Stata to create the histogram for observations where the value of *gender* is 1. Unfortunately, the result is that one histogram was overlaid on top of the other, making the comparison very difficult, since we can't see all aspects of both histograms. Again, Stata has an option to make our life easier. Instead of plotting the histogram, we can tell Stata to plot the kernel density estimates. All you need to know is that it is a graph that represents the histogram only using a line instead of the bars:

twoway (kdensity engage if gender ==0) (kdensity engage if gender == 1)



We now have a much better graph that we can use to compare. But which of the graphs is for males and which is for females? The legend doesn't really help us. In our command, we specified the gender == 0 condition first, and this represent males, so the first graph in the legend (the blue one) is for males. It would be better to have output that reflects this. This is accomplished by using the **legend()** option. This option allows us to control the text that appears in the legend, as well as many other things:

```
twoway (kdensity engage if gender ==0 ) (kdensity engage if gender == 1),
legend(label(1 "Males") label(2 "Females"))
```

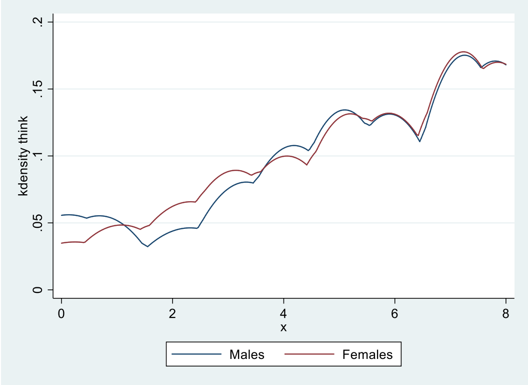


The command is longer than what we have been using so far but it’s actually simple. Using the **legend()** option, we told Stata that it should label group 1 as “Males” and group 2 as “Females”.

We are now capable of comparing the two distributions. What we notice is that the plot for females leans more towards the left than that of males. This indicates that females report lower levels of engagement in academic misconduct than males.

Let us produce the same graph but this time for the variable *think*, which measures to what extent the student classifies certain acts as cheating:

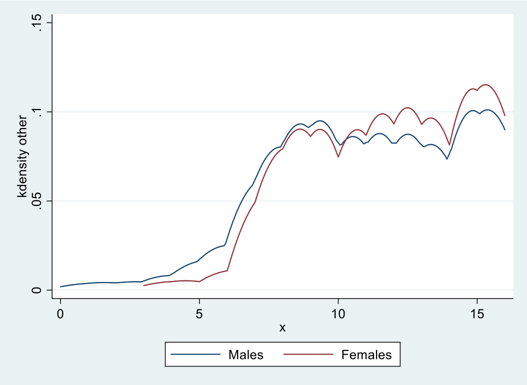
```
twoway (kdensity think if gender ==0 ) (kdensity think if gender == 1), legend(label(1
"Males") label(2 "Females"))
```



We used the exact same command but instead of plotting *engage* we plot *think*. Unlike in the previous case, we don’t see much difference between males and females. However, the shape of the distribution of *think* is very different from the shape of the distribution of the variable *engage*. For *engage*, we see that most students reported average levels of engagement. For *think* on the other hand, we saw that more students reported higher levels.

Let us now do the same for the variable *other*:

```
twoway (kdensity other if gender ==0 ) (kdensity other if gender == 1), legend(label(1
"Males") label(2 "Females"))
```



This is a very interesting graph, especially when we compare it to the graphs for the variable *engage*. While the largest density for *engage* was around the middle, the largest density for *other* is at the far right. What this indicates is that students believe that other students cheat a lot while they report that they sometimes cheat. The best way to understand this is to compare these two variables by using the same plot. However, this will be the goal of a future section. For now, we are only plotting the same variable but for different groups on the same plot, but as you can see, there is a picture that is starting to emerge. Females have higher GPAs, and they cheat less. Students report average levels of engaging in cheating but believe that other students engage a lot in cheating.