

## 12 Professor Salary Lab Project

The data that we are working with contains information about the salaries and numbers of faculty members at colleges and universities across the United States (including here at UCLA). To get the data into *Stata*, type the command:

```
. use http://www.stat.ucla.edu/labs/datasets/profsal.dta
```

With this lab, many questions you might have had about the pay scale of your professors can be answered. Maybe you have wondered how much the salary for a full professor influences the number of full professors at a school. Maybe you've wondered if the number of associate professors at a school has any relationship to the number of assistant professors. These are both questions that can be answered with regression analysis. What do preliminary scatterplots tell us about these questions?

Maybe you're interested in how much more full professors make than associate professors. To do this we'd have to generate a new variable:

```
. generate saldif=avesalfull-avesalaso
```

Then a boxplot would be nice to illustrate the range of this difference. What does the histogram look like? To get a nice histogram, the `graph` command with the `xline` and `bin` options can be used.

```
. help graph
```

Using this help menu, create a histogram with 20 bins and a reference line at  $x = 0$  (zero). What is interesting about this histogram (see Figure ??)?

We can also look at how two states in different regions of the country compare. As an example, we'll look here at how Illinois and Georgia compare in certain ways.

To eliminate all observations that are from states other than Illinois and Georgia, the “keep” command is used. For more information on the format of the “keep” command, type

```
. help keep
```

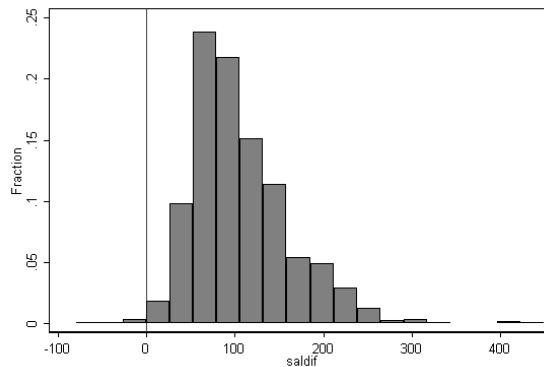


Figure 1: Histogram of `saldif` with  $x = 0$  reference line

So, in this example we would type

```
. keep if state=="IL" | state=="GA"
```

into our command line since these are the only states we are interested in. Note that the “|” mark is a pipe, not a number one or the letter “l”.

Once we have only the data of interest in our dataset we can type

```
. by state: summarize
```

to get a better idea of what the data looks like. Notice that the summary reports that there are no observations present for the first three variables; `state`, `type`, and `schoolname`. This is because these are character rather than numerical variables and no numerical summaries of them are possible. Don’t worry, the data is there.

Now consider in what ways you’d like to compare colleges in the two states. Maybe you’re interested in the size of schools and see the number of faculty at an institution as a good measurement of this. To do side by side boxplots of the `numfaculty` variable as shown in Figure ?? for this data, type in the command:

```
. graph numfaculty, box by(state)
```

What do we see here about school sizes in Georgia (the box on the left) compared to school size in Illinois?

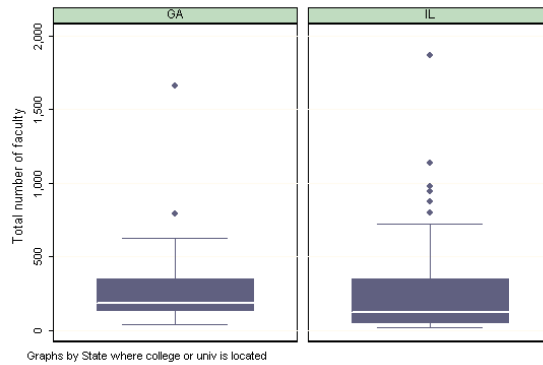


Figure 2: Side by Side Boxplots of numfaculty for Georgia and Illinois

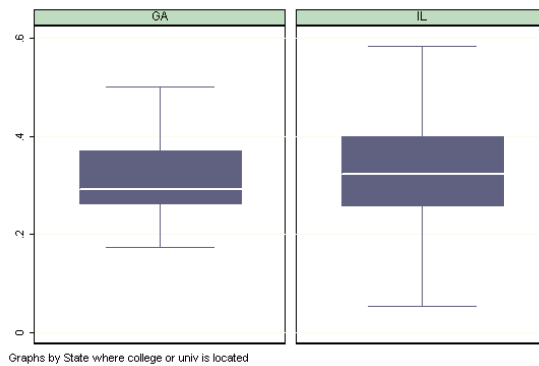


Figure 3: Side by Side Boxplots of pctfull for Georgia and Illinois

Maybe you're more interested in what percent of faculty are full professors. To answer this question a new variable will need to be generated. Using the "gen" command we've seen in earlier labs (or the equivalent "generate" command) we can create a variable that shows the percent of faculty that hold the rank of full professor at each school.

```
. generate pctfull=numfull/numfaculty
```

And then side by side boxplots by state can be created similarly to how we created the ones for number of faculty. The results for this example are shown in Figure ??.

What is interesting about these box plots? What else might be interesting to

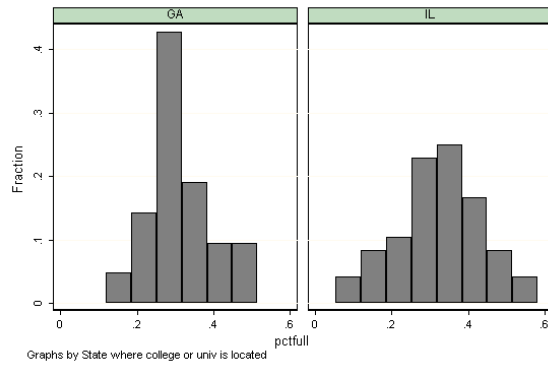


Figure 4: Side by Side Histograms of pctfull for Georgia and Illinois

see? Side by side histograms of the pctfull variable are shown in Figure ?? . Remember that the “bin()” option allows you to generate histograms with more than the default five groups.

## Assignment

This lab handout has mentioned only a few of the many questions that can be asked about this data set and shown you only a few examples of applicable commands you can use to answer those questions. Your job is to come up with questions of your own, answer them with *Stata*, and type up a report describing your findings. You should consider questions about the entire dataset, as well as questions about how two or more states compare (choose states other than Illinois and Georgia). Remember that the “help” and “search” commands can be very useful. Your report should be at least three pages in length, with all graphs and/or charts you reference attached at the end.