Stata Walkthrough: Exploring and Describing Data

This is an optional assignment, intended to get you accustomed to working with data in Stata. To do this assignment, you will need to go to my webpage and download the database of SAT scores by U.S. states.

1. Open the database using the menus (or open it directly from your computer).

2. Type `describe` to look at the types of variables and their labels. `float` are normal, continuous values; `int` and `byte` take discrete numerical values; `str` have text. You will also see the labels attached to each variable, which should explain what they measure.

3. Type `list` to view all of the observations in your database. You will see the state's name, poverty rate, percentage of students taking the SAT, scores on the test, and per student spending for each state.

4. Type `summarize` to get some important summary statistics for each state (the mean, standard deviation, minimum and maximum). Note that you have zero (numerical) values for the state name, so Stata does not give you summary statistics for this variable.

5. Type `summ spend, detail` to get detailed summary statistics for this variable. In the “Percentiles” column, you determine the lower and upper quartiles (25th percentile is $6409; 75$th percentile is $8700.50$) as well as the median (50th percentile is $7080.50$). The mean, standard deviation, and variance are also shown.

6. You can use Stata as a calculator by typing “display” followed by some expression. (The commands for basic operations are standard: +, -, *, /, and ^ for exponents. Stata understands parentheses, and knows to always Please Excuse My Dear Aunt Sally.) To confirm that the standard deviation is the square root of the variance, type `display (1205623)^(0.5)`.

7. You might be curious if spending is correlated with some other variables. Type `corr spend pov percent sat` to obtain a table of all the correlation coefficients between the variables. In the first column, you can tell that there is a moderate negative correlation between spending and poverty rates (states with higher poverty rates tend to spend less per student), a moderate positive correlation between spending and the percentage who take the test (states that spend more tend to have more students who take the exam), and (surprisingly) a small negative correlation between spending and SAT scores. The only explanation I could come up with was that higher spending encourages the more marginal students to take the exam. (Better teachers might encourage more students to think about college and take the exam, but this isn’t going to have any effect on the types of students who score 1600s.) Sure enough, if you check the correlation between the number of students taking the exam and the average score (the third column of this table), you find a very strong negative correlation.
8. You might want to look at a scatterplot of the relationship between spending and poverty rates. The command is `scatter spend pov`. Maybe I see a slight negative relationship, but it’s far from clear. This is consistent with a moderate correlation coefficient.

9. Now you might want to divide your database into categories based on poverty rates—specifically, the least impoverished 25% of states, the next 25%, then the next 25%, and then the 25% most impoverished. Type `sum pover, detail` to determine the values of these quartiles: 9.4% at the 25th percentile, 11.2% at the 50th percentile, and 14.1% at the 75th percentile. A “dummy variable” is something that takes a value of one if an observation falls into a particular group, and zero if not. It’s a useful way to code qualitative data. (For example, you could convert sex, which is “male” or “female” into a dummy variable called `male`, which takes a value of one for men and zero for women.) Construct dummy variables this way:

   ```
   gen p1 = (pover <= 9.4)
   gen p2 = (pover > 9.4 & pov <= 11.2)
   gen p3 = (pover > 11.2 & pov <= 14.1)
   gen p4 = (pover > 14.1)
   ```

10. You can also create a variable recording the poverty quartile of the state. One way to do this is:

   ```
   gen povquart = 1 if p1==1
   replace povquart = 2 if p2==1
   replace povquart = 3 if p3==1
   replace povquart = 4 if p4==1
   ```

   The first command gives the variable a value of 1 if the condition “p1==1” is met. (Remember that Stata requires a double-equal when you’re talking about a condition; a single-equal is reserved for formulas.) If the condition is not true, Stata gives the variable a missing value, since it’s uncertain what it should be. The following steps replace those missing values with 2 if the state is in the second quartile, a 3 if the state is in the third quartile, and a 4 if it is in the fourth quartile.

11. You might want to label these variables in order to remember what they are. You can type `label variable povquart “Poverty quartile”`.

12. Now look at the distribution of the variable by typing `tab povquart`. You should find that approximately 25% of the database is in each quartile. (“Approximately” because you can’t have exactly 12.5 states in each group.)

13. You might want to see how spending varies between the poverty quartiles. Type `graph box spend, over(povquart)` to produce this graphic. You should be able to see that the median spending drops from one group to the next (except maybe between the 3rd and 4th poorest groups). There is probably not any clear association between the poverty quartile and the range of spending, however.
14. Are the medians of the third and fourth poorest states the same? You can obtain these by typing:

   ```
   summ spend if povquart==3, detail
   summ spend if povquart==4, detail
   ```

   The third is $6762.50 and the fourth is $6772, so the median spending of the most impoverished group of states is actually a bit higher. (However, you can see that the average is lower: $6973.85 versus $7023.74.)

15. In the side-by-side box plots, you might have noticed a few outliers. The most peculiar one is the state in the least impoverished quartile that spends a very small amount per student (under $5000). Which state is this? You can type `browse` to view the data and search for it. If you want to make things a bit easier, type `sort spend` and then `browse`, and the lowest spending states will appear at the top. Utah is the weird state.

16. There's a bit of a problem in the data, which probably exaggerates the negative relationship between spending and poverty rates: poverty rates tend to be higher in the South and West than they are in the Northeast and Midwest. At the same time, the cost of living (as well as the cost of hiring teachers, etc.) is substantially lower in the South and West. Type `sort pov` and then `browse` to see the first relationship. (The least impoverished states are New Hampshire, Minnesota, Connecticut, and New Jersey; the most impoverished states are Mississippi, Louisiana, New Mexico, and West Virginia.) Then type `sort spend` and `browse` to see the states that spend the most (New Jersey, North Dakota, New York, and Connecticut) and the ones that spend the least (Utah, Arkansas, California, and Idaho). This isn't proof, but it's clear there's a regional association between the variables. This is a common problem with many poverty statistics.

17. What about SAT scores for the different groups? When you type `graph box sat, over(povquart)`, you probably don't see any clear relationship. Surprisingly, the 3rd quartile (the next-to-the-most impoverished states) have the highest median SAT scores. Is the relationship the same if we look at only verbal and math scores? `graph box verb, over(povquart)` and `graph box math, over(povquart)` will tell us the answer. (Basically, yes.)

18. Another way to compare our groups is to create a table with the means and standard deviations for each: `tab povquart, summ(sat)`. This paints a slightly different picture: SAT scores tend to be declining as poverty rates are higher, except for that funny third quartile. We can also see that the third quartile has the highest variance (or standard deviation, really) in outcomes.

19. Finally, I wonder if there's one oddball in this third quartile that is driving the result. You can type `list state pov sat if povquart==3` in order to get a description of these three variables for states in this group. In fact, there is: while most SAT scores are in the range from upper-900s to
mid-1100s, there’s a single state with an average score well above the group—North Dakota has an average SAT score of 1207. You can type in `summ sat` to see that 1207 is in fact the highest average SAT score of any state in our database.

20. You might want to repeat the table of SAT scores over poverty, omitting North Dakota from the calculations. The command to do this would be `graph box sat if state != "N. Dakota", over(povquart)`.
(The exclamation mark is geek-speak for “not”, so “!=” means “not equals”.) But no, not really—the third quartile still has the highest median SAT scores.