Recitation 2: Variable types, Storage types, and Value types

McCourt School of Public Policy, Georgetown University

Key Ideas:

- Categorical variables
- Reporting results
- Help files

Log Files for Recording a Session [CHANGE NEEDED]

- Last week we used a log to record the final output from a do-file.
- You can also use them to keep a record of an interactive session.
- Useful in recitation, so you have a record of the commands that you used.
- Start a log now, to record commands and output from this recitation.

Categorical Variables

- Numeric variable, but values has no intrinsic meaning
- A type of discrete variable in that it takes a countable number of values.
- But numeric values are not meaningful as numbers.

For example,

1. let's look at the industry variable in a dataset.

   ```
   clear

   sysuse nlsw88.dta
   ```

2. Browsing the data, we see BLUE text that indicates labels - click on the observations and you see that each industry is assigned a number. For example, `agriculture/fisheries` is coded as `1`. This type of data is a `CATEGORICAL` variable.

   ```
   summarize industry
   ```

3. Results from summarize are not meaningful - we can't have a mean or standard deviation for categories! (Don't report mean and st.dev. for categorical variables)

4. We can learn more about the industry variable with `tabulate` (abbrv. `tab`)

    ```
    tabulate industry

    tabulate industry, nolabel
    ```

5. Tabulate shows counts of observations - we can also get that info using the command `count`

    ```
    count if industry==1

    count if industry==2

    count if industry==3

    count if industry==4
    ```

6. Tabulate gives us the distribution of the industry variable. The same data can be displayed graphically:

    ```
    histogram industry, discrete
    ```

7. If you want to see tabulate and label information in one place, `codebook` is useful.

    ```
    codebook industry, tabulate(12)
    ```

8. Summary statistics can be reported for each category - you have to stratify or segment your categories to get useful stats.

    ```
    summarize wage hours if industry==1

    summarize wage hours if industry==2

    summarize wage hours if industry==3

    summarize wage hours if industry==4
    ```

    - Or, a faster way:

        ```
        tabstat wage hours, statistics(mean sd count) by(industry)
        ```

We will spend more time with tabstat later

Binary Variables

- There is one type of categorical variable for which the numeric value has meaning
- Binary variables describe a non-numeric characteristic of the observation, such as gender, race, etc.
- A binary variable equals 1 if the observation has that characteristic and zero if it does not.

Example: married

- Browsing through the data, we see…
- for married people, the value of the `married` variable is `1`
- for non-married people, aka single people, the value of the `married` variable is `0`
- We can use binary variables just like other categorical variables:

```
tabulate married tabulate married, nolabel count if married==1
count if married==0

codebook married

summarize wage hours if married==0 summarize wage hours
if married==1 tabstat wage hours, statistics(mean sd count)
by(married)
```

- But, the numeric value is ALSO meaningful with binary variables!

```
summarize married
```

- 64.2% of the people in this data set are married.
- To see this another way:

```
tab married display 1442/2246
```

*** Sample Questions *** Create a do-file to answer the questions below. Document questions and answers with comments.

1. Report appropriate summary statistics for the following variables: age, race, grade, collgrad, union
2. What is the most common industry for workers in this sample?

3. What is the average wage for that industry?

   clear

   sysuse nlsw88.dta

4. Report appropriate summary statistics for the following variables: age, race, grade, collgrad, union

   summarize age grade collgrad union

- We can't summarize race - mean and sd don't mean anything!

  tabulate race tabulate collgrad tabulate union * We can tabulate the categorical and binary variables. 2. What is the most common industry for workers in this sample?

```
>tabulate industry
```

Ans: Professional Services 3. What is the average hourly wage for that industry?

  tabulate industry, nolabel summarize wage if industry==11

Ans: $7.87 per hour

Reporting Results

Professors have different requirements for reporting answers. Some may not accept copy/paste Stata output. But you still may want to use copy/paste for your own notes.

Two ways: 1. copy text, paste to Word - change font to courier new - change size to 8, 9, or 10 - change line spacing to 0 pt. 2. copy table, paste to Excel - format in Excel (use formatting wizard)

Help Files

- We used several complicated commands today with options and if statements.
- You do not have to memorize these commands.
- Every command has a help page to tell you how to use it.

  help count

- Overview of help page, including `Also See`
- Explain syntax elements relevant to count command:

  - **bold** - type as is
  - ***italics*** - replace with your own variables or expression
  - [brackets] - optional
  - underline - shortest abbreviation
  - blue - hyperlinks

- Use `Also See` to go to help page for tabulate oneway
- Explain new syntax elements: varname, options (skip **`weights`**)
- For more in depth description and examples, go to manual entry with `Also See`
- Look at another help page: tabstat

  help tabstat

- Explain varname vs. varlist
- Go through examples on tabstat help page
- Most help pages are easy to find by guessing the command name and `See Also`
- But there are a few that are worth remembering:

  1. help contents_should_know
  2. help language
  3. help operators

- If you forget these, they can all be found within a few clicks using menus: Help Contents Basics . . .

Log File

- Close log file and look at output before moving on to after class exercise.

After Class Exercise

Answer the questions below using the census.dta data sets. Create a do-file and log-file showing your work. Document answers using comments.

1. The `region` variable assigns US states to different regions of the country. How many different regions are used in this data?
2. What are the names of the different regions?
3. Which region has the largest number of states?
4. Write a command to tabulate the regions in descending order of number of states.
5. Write a command to list the states in the largest region.
6. Which region has the highest average median age? What is the median age of that region?
7. Are there any string variables in this data set? If so, which variables?
8. Use tabstat to show the total number of marriages and divorces for each region.

After Class Exercise Answer

```
clear
sysuse census.dta
```

1. The `region` variable assigns US states to different regions of the country. How many different regions are used in this data?

   tabulate region Ans: 4 regions

2. What are the names of the different regions?

   Ans: NE, N Cntrl, South, West

3. Which region has the largest number of states?

   Ans: South

4. Write a command to tabulate the regions in descending order of number of states.

   tabulate region, sort

5. List the states in the largest region.

   tabulate region, nolabel codebook region list state if region==3

6. Which region has the highest average median age? What is the median age of that region?

tabstat medage, by(region) Ans: NE , 31.2333 years

7. Are there any string variables in this data set? If so, which variables?

   describe Ans: Yes, state and state2

8. Use tabstat to show the total number of marriages and divorces for each region.

   tabstat marriage divorce, statistics(sum) by(region)