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

McCourt School of Public Policy, Georgetown University

Key Ideas:

- String vs. Numeric Variables
- Missing Values
- Categorical variables
- Reporting results

Variables Types: String vs. Numeric

Strings are stored as text, and cannot be added, subtracted, etc.

Example variable: `make`

- browse dataset
- Contrast string variable `make` with labeled numeric variable, `foreign`
  - Example: Generate tables of frequencies with/without labels with ‘, nolabel’ on certain commands

    ```
    tabulate foreign
    tabulate foreign, nolabel
    tabulate make
    tabulate make, nolabel
    ```

  - Strings are red, labeled numeric are blue
  - For strings, the actual value is text.
  - For labeled numeric, the actual value is a number

```
sysuse auto, clear
describe
```

- Look at storage type column
- Anything that begins with `str` is a string
- Anything else is a number, including byte, int, float, long, double.

```
summarize
```

- String variables are not included in summary statistics, because it is not possible to perform calculations on them.
- We will cover string variables in more depth.
- For now, just know what they are and how to identify them.

Missing Values

Sometimes not every observation has a value for every variable. In survey data, this could arise if people skip a question.

We will talk more about missing data in Quant 2. For now, you should just know that missing data exists and what it looks like in Stata.

In Excel, missing data is represented simply as a blank cell. In Stata, missing data for numeric variables is represented by a period or dot `.` In Stata, missing data for string variables is represented by an empty string ‘

Missing data is not included in calculations for summary statistics

- Example: `rep78`
- There are 74 observations, or rows, in this data set -Enter `describe` or `count`
- But rep78 only has 69 observations in summary statistics -Enter `summarize rep78`
- The difference is because of observations with missing data are not included in the calculation. `-display 74-69`
- 5 observations with missing data
- `browse`
- Observations with a variable missing data have a `.` representing this when viewing them in the browser.
- We can see this with the count command `count if rep78==.`
- For string variables, missing data appears blank. Try `count if make==""`

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,

Let's look at the industry variable in a dataset.

```
clear
sysuse nlsw88.dta
```

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.

```
* WRONG COMMAND
summarize industry
```

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)

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

```
tabulate industry
tab industry, nolabel
```

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
```

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

```
histogram industry, discrete
```

This graph can also be made by going to Graphics -> Histogram

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

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

Codebook * For numeric data, it reports the range, percentiles, and missing values. * For categorical data just calling codebook provides an example of the use * When calling ‘codebook variable, tabulate(number)’ , codebook lists the frequency and variable name and label for the number of unique values called for * If you do not know how many values there are, using a large number in tabulate lists all unique values.

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
```

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
  ```

- 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
```

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)

Log File

- Stata can record your session into a file colled a log file but does not start a log automatically; you must tell Stata to record your session.
- There are two types of log files:
  - `.log` is a text file that can be opened with many programs, Word, Notepad, etc.(Recommended)
  - `.smcl` is a special Stata format that can only be opened with Stata. `.smcl` formats output just as it appears in results window, while `.log` is plain text.
  - If you are submitting a log file with an assignment, always use `.log`

Logs begin recording when they are opened/started, and stop recording when they are closed. Start recording a log file using the drop-down menu (File > Log > Begin…). Stop recording a log file using the drop-down menu (File > Log > Close).

In Class Activity 1

Create a do-file with `nlsw88.dta` to answer the questions below.

Create a `do-file` and `log-file` showing your work.

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 hourly `wage` for that `industry`?

```
clear
sysuse nlsw88.dta
* 1. 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
*or
codebook industry, tabulate(99)

* 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
```

In Class Activity 2

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?

In Class Activity 2 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?

   ```
   sum medage if region==1
   sum medage if region==2
   sum medage if region==3
   sum medage if region==4
   ```

   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