# Introduction to Geographic Information Science

Week 01 Stata Commands

Christopher G. Prener, Ph.D.

Spring, 2016

## Contents

# 1 Using Pre-loaded Stata Datasets

## 1.1 Syntax

```
sysuse ["]file_name["]
sysuse dir
```

## 1.2 Examples

To load a dataset that comes pre-loaded with Stata:

```
. sysuse census.dta
(1980 Census data by state)
```

To generate a list of all pre-loaded datasets:

```
. sysuse dir
```

| | | | |
|---|---|---|---|
| auto.dta | citytemp.dta | nlsw88.dta | tsline2.dta |
| auto2.dta | citytemp4.dta | nlswide1.dta | uslifeexp.dta |
| autornd.dta | educ99gdp.dta | pop2000.dta | uslifeexp2.dta |
| bplong.dta | gnp96.dta | sandstone.dta | voter.dta |
| bpwide.dta | lifeexp.dta | sp500.dta | xtline1.dta |
| cancer.dta | network1.dta | surface.dta | |
| census.dta | network1a.dta | tsline1.dta | |

## 1.3 Notes

I recommend saving a copy of the pre-loaded datasets by using the `save` command before making changes to them. Additional details can be found in the Stata documentation for the `sysuse` command (link).

# 2 Describing Data

## 2.1 Syntax

```
describe [varlist] [, options]
```

## 2.2 Examples

To list all of the variable names and labels of a loaded dataset:

```
. sysuse census.dta
(1980 Census data by state)

. describe

Contains data from /Applications/Stata/ado/base/c/census.dta
  obs:            50                          1980 Census data by state
 vars:            13                          20 Jan 2016 15:53
 size:         2,900
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
state           str14   %-14s                 State
state2          str2    %-2s                  Two-letter state abbreviation
region          int     %-8.0g     cenreg     Census region
pop             long    %12.0gc               Population
poplt5          long    %12.0gc               Pop, < 5 year
pop5_17         long    %12.0gc               Pop, 5 to 17 years
pop18p          long    %12.0gc               Pop, 18 and older
pop65p          long    %12.0gc               Pop, 65 and older
popurban        long    %12.0gc               Urban population
medage          float   %9.2f                 Median age
death           long    %12.0gc               Number of deaths
marriage        long    %12.0gc               Number of marriages
divorce         long    %12.0gc               Number of divorces
-------------------------------------------------------------------------------
Sorted by:
```

To list only a selection of a dataset's variables:

```
. describe state pop death

              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
state           str14   %-14s                 State
pop             long    %12.0gc               Population
death           long    %12.0gc               Number of deaths
```

### 2.3 Options

The `fullnames` option forces Stata to display even long variable names. By default, long variable names will be truncated. While it is good practice to limit the length of variable names, you may receive data from a third party that has not been curated with care. In that case, the `fullnames` option may be required until you can clean the data properly.

### 2.4 Notes

Additional details can be found in the Stata documentation for the `describe` command (link).

# 3 Basic Descriptive Statistics

## 3.1 Syntax

```
summarize [varlist] [, options]
```

## 3.2 Example

To generate descriptive statistics for a variable:

```
. sysuse census.dta
(1980 Census data by state)

. summarize pop

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         pop |         50     4518149     4715038     401851   2.37e+07
```

The `Obs` refers to the number of valid observations in the datasets. Think of this as the number of rows that contain data for the variable `pop`. The `Mean` refers to the average. In 1980, the average population of a U.S. state was 4,518,149 people.

The `Std. Dev.` refers to a measure of central tendency known as the standard deviation. Assuming the population of U.S. states in 1980 was normally distributed (i.e. follows what is commonly called the bell curve), we could estimate that two-thirds of states have populations that fall between 0 and 9,233,187 people. We found these two numbers by subtracting the standard deviation from the mean for the lower bound, and adding the standard deviation to the mean for the upper bound. Technically the lower bound of this calculation is -196,888, but we say that our estimate is 0 because it is not possible for a state to have a negative population.

The `min` refers to the smallest single value and the `max` refers to the largest single value. Subtracting the minimum value from the maximum value gives us the range of the variable's distribution.

### 3.3 Options

To generate more extensive descriptive statistics, use the `detail` option:

```
. sysuse census.dta
(1980 Census data by state)

. summarize pop, detail
```

```
                        Population
-------------------------------------------------------------
      Percentiles      Smallest
 1%       401851         401851
 5%       511456         469557
10%     671742.5         511456       Obs                  50
25%      1124660         594338       Sum of Wgt.          50

50%      3066433                      Mean            4518149
                        Largest       Std. Dev.       4715038
75%      5463105       1.19e+07
90%     1.11e+07       1.42e+07      Variance        2.22e+13
95%     1.42e+07       1.76e+07      Skewness        2.073804
99%     2.37e+07       2.37e+07      Kurtosis        7.729186
```

The `detail` option gives percentile values for the distribution of the variable. Remember that the 50th percentile is also the variable's median - in this case, half of all states have a population less than 3,066,433 and half have a population greater than that value. The next column contains the smallest four values and the largest four values in the distribution. Finally, on the right side of the output, you are given the variance, skewness, and kurtosis of the distribution.

For large numbers, values will be reported using scientific notation. In order to convert that to a more intuitive value, use the `display` command:

```
. display 1.11e+07
11100000
```

We can therefore see that the 90th percentile of population is 11,100,000 individuals.

# 4 Common Errors

## 4.1 Error 111: Variable Not Found

When referencing a variable using the `describe` or `summarize` commands, you may receive this error:

```
. describe type
variable type not found
r(111);
```

This error indicates one of two conditions: (1) the variable does not exist in the dataset or (2) there is a misspelling in the variable name.

## 4.2 Error 198: Option Not Allowed

If an option is misspelled or is not actually a valid option for a command, you will see the following error. In this case, the option for the `summarize` command has been misspelled - it is `detail`, not `detailed`:

```
. summarize pop, detailed
option detailed not allowed
r(198);
```

## 4.3 Error 199: Command Not Recognized

If a command is misspelled, it will generate the following error. In this case, the command `describe` has been misspelled:

```
. dscribe type
command dscribe is unrecognized
r(199);
```

If the command was written as part of a user-installed package, this error could also indicate that package has not been installed locally.

## 4.4 Error 601: Data Not Found

When referencing a file using the `sysuse` command, you may receive this error:

```
. sysuse data
file "data.dta" not found
r(601);
```

This error indicates one of two conditions: (1) the file is not one that comes pre-loaded with Stata or (2) there is a misspelling in the file name.