FUNDAMENTALS OF DATA ANALYSIS USING STATA

Course detail: https://julhas.com/jsedutech/stata-level-one.html
Mentor: Julhas Sujan

Stata is a general-purpose statistical software package created in 1985 by StataCorp. Most of its users work in research, especially in the fields of economics, sociology, political science, biomedicine, and epidemiology.

Session agenda:

- Stata download and installation
- Dummy excel dataset and overview
- Basic data management in excel
- Excel dataset import
- Data management in Stata
    - Create new variable
    - Drop variable
    - Rename
    - Recode
- DO and Log file
- Append and Merge
- Frequency distribution
- One way and cross table
- Measures of central tendency
- Different statistical test
- Graphics
- Linear regression
- Logistics regression

1. Stata download and Installation

- Download Stata from here: https://julhas.com/jsedutech/materials/STATA16.zip
- Installation guideline: https://julhas.com/jsedutech/materials/Level-1/Stata-Session-1.pdf
- Stata window overview: https://julhas.com/jsedutech/materials/Level-1/Stata-Session-2.pdf

2. Excel dataset, variable overview and basic data management

- Download dataset: https://julhas.com/jsedutech/materials/TB-Dummy-Dataset-for-JS-Edutech.xlsx

3. Excel dataset import

- Download session materials: https://julhas.com/jsedutech/materials/Level-1/Stata-Session-3.pdf

4. Data management in Stata

- Browse dataset
  br

- Directory and detail
  dir

- Current date
  display c(current_date)

- create some observations – still no variables
  set obs 5

- create a variable named x, which has the
- value of 1 for all observations
  generate x = 1

- create another variable y, which has the
- observation number as its value
  generate y = _n

```
list
```

- Variable detail check

  ```
  codebook gender
  ```

- Describe variable

  ```
  describe age
  ```

- Create new variable

  ```
  gen testVariable
  ```

- Drop/ delete variable

  ```
  drop testVariable
  ```

- Drop missing values

  ```
  drop if variableName == .
  ```

- Variable leveling

  ```
  label variable bmi "Body Mass Index"
  ```

- List of data of a variable

  ```
  list age gender
  ```

- Rename

  ```
  gen genderTest = gender
  rename genderTest Sex
  ```

- Encoding:

  ```
  encode gender, gen(genderBinary)
  encode division, gen(divisionBinary)
  encode siteofdisease, gen(siteofdiseaseBinary)
  encode drugresistancetype, gen(drugresistancetypeBinary)
  encode currentregimen, gen(currentregimenBinary)
  ```

- Recoding

  ```
  gen ageGroup = age
  recode ageGroup (0/4 = 0) (5/14 = 1) (15/24 = 2) (25/34 = 3) (35/44 = 4) (45/54 = 5) (55/64 = 6) (65/120 = 7)
  label define ageGroup 0 "<=4 Years" 1 "5-14 Years" 2 "15-24 Years" 3 "25-34 Years" 4 "35-44 Years" 5 "45-54 Years" 6 "55-64 Years" 7 ">= 65 Years"
  label values ageGroup ageGroup
  tab ageGroup
  ```

- Replace data of a variable

  ```
  replace address = "Dhaka" if sn > 10
  ```

- Missing data check
  ```
  count if age =.
  ```

5. DO and Log file

- Check detail: https://julhas.com/jsedutech/materials/Level-1/Stata-Session-3.pdf (Page -9)

6. Append and Merge

- Session detail: https://julhas.com/jsedutech/materials/Level-1/Stata-Session-6.pdf
- Append:
  ```
  use "dataset-1.dta"
  append using "dataset-2.dta"
  ```

- Merge two datasets
  ```
  use "dataset-1.dta"
  merge 1:1 sn using "dataset-2.dta"
  ```

7. Frequency distribution

- Table of gender variable
  ```
  tab gender
  ```

- Sorting
  ```
  bysort age: tab gender
  ```

- Find average
  ```
  egen average  = rmean (var1 var2 var3)
  ```

- Describe all variables
  ```
  describe
  ```

- Detail of all variables
  ```
  codebook
  ```

- Unique and missing obs check
  ```
  inspect
  ```

- Summarize
  ```
  summarize age
  summarize bmi, detail
  ```

- Data sorting
  ```
  sort sn
  by age: summarize bmi
  tab age, sort
  ```

- tabulate command is useful for obtaining frequency tables
  ```
  tab gender
  tabulate gender ageGroup
  ```

- The tab1 command can be used as a shortcut to request tables for a series of variables
  ```
  tab1 gender age bmi
  ```

- use the plot option to make a plot to visually show the tabulated values.
  ```
  tabulate gender, plot
  ```

- Tabulate and summarize
  ```
  tabulate age, summarize(bmi)
  ```

8. One way and cross table

- One-way table of frequencies for v1
  ```
  tabulate v1
  ```

- Sort table in descending order of frequency
  ```
  tabulate v1, sort
  ```

- Generate indicator variables v1 1, v1 2, . . . representing the levels of v1
  ```
  tabulate v1, generate(v1_)
  ```

- Treat missing values like other values of v1
  ```
  tabulate v1, missing
  ```

- Display numeric values of v1 rather than value labels
  ```
  tabulate v1, nolabel
  ```

- Create one-way tables for v1, v2, and v3
  ```
  tab1 v1 v2 v3
  ```

- Cross table
  ```
  tabulate ageGroup gender
  ```

- Cross table with percentage
  ```
  tabulate ageGroup gender, column
  tab ageGroup gender, col
  ```

- nofreq option to suppress the frequencies, and just focus on the percentages
  ```
  tabulate ageGroup gender, column nofreq
  ```

- Column change
  tabulate ageGroup gender, nofreq column

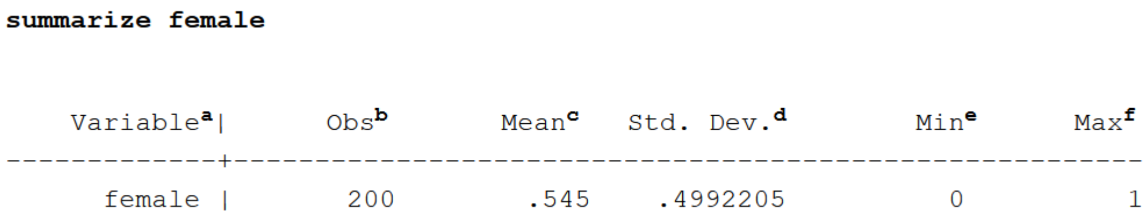
9. Measures of central tendency

- Detail: https://julhas.com/jsedutech/materials/Level-1/Stata-Session-10.pdf

Mean and Median for Age

tabstat age, stats(mean median)

The summarize command, which also can be abbreviated to sum, gives a little less information, and is best used on continous variables where the mean is of interest



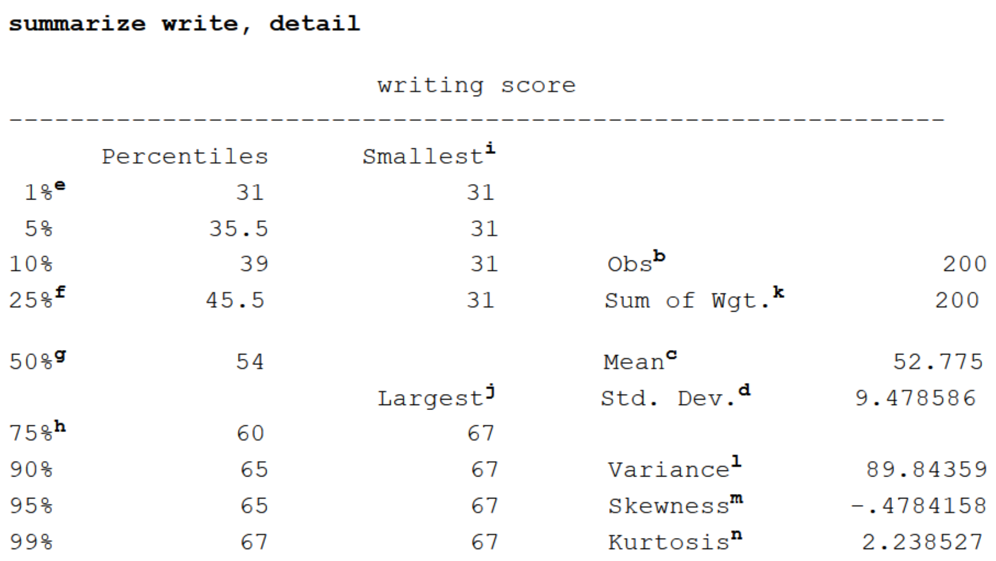
```
summarize female

    Variable[a]|       Obs[b]        Mean[c]    Std. Dev.[d]       Min[e]        Max[f]
-------------+--------------------------------------------------------
      female |       200        .545     .4992205          0          1
```

a. Variable – This column indicates which variable is being described. You can list more than one variable after the summarize command; when you do, you will see each variable on its own line of the output.

b. Obs – This column tells you the number of observations (or cases) that were valid (i.e., not missing) for that variable. If you had 200 observations in your data set, but you had 10 missing values for the variable female, then the number in this column would be 190.

c. Mean – This is the mean of the variable. In this case, our variable female ranges from 0 to 1 (the min and max values), so the mean is actually the proportion of observations coded as 1.

d. Std. Dev. – This is the standard deviation of the variable. This gives information regarding the spread of the distribution of the variable.



```
summarize write, detail

                           writing score
-------------------------------------------------------------
      Percentiles      Smallest[i]
 1%[e]        31             31
 5%         35.5             31
10%           39             31       Obs[b]                 200
25%[f]      45.5             31       Sum of Wgt.[k]         200

50%[g]        54                      Mean[c]             52.775
                        Largest[j]    Std. Dev.[d]      9.478586
75%[h]        60             67
90%           65             67       Variance[l]       89.84359
95%           65             67       Skewness[m]      -.4784158
99%           67             67       Kurtosis[n]       2.238527
```

e. 1% – This is the first percentile. Percentiles are calculated by ordering the values of a variable from lowest to highest, and then finding the value that corresponds to whatever percent you are interested in, in this case, 1%. Hence, 1% of the values of the variable write are equal to or less than 31.

f. 25% – This is the 25th percentile, also known as the first quartile.

g. 50% – This is the 50th percentile, also known as the median. If you order the values of the variable from lowest to highest, the median would be the value exactly in the middle. In other words, half of the values would be below the median, and half would be above. This is a good measure of central tendency if the variable has outliers.

h. 75% – This is the 75th percentile, also known as the third quartile.

i. Smallest – This is a list of the four smallest values of the variable. In this example, the four smallest values are all 31.

j. Largest – This is a list of the four largest values of the variable. In this example, the four largest values are all 67.

k. Sum of Wgt. – This is the sum of the weights. In Stata, you can use different kinds of weights on your data. By default, each case (i.e., subject) is given a weight of 1. When this default is used, the sum of the weights will equal the number of observations.

l. Variance – This is the standard deviation squared (i.e., raised to the second power). It is also a measure of spread of the distribution.

m. Skewness – Skewness measures the degree and direction of asymmetry. A symmetric distribution such as a normal distribution has a skewness of 0, and a distribution that is skewed to the left, e.g., when the mean is less than the median, has a negative skewness.

n. Kurtosis – Kurtosis is a measure of the heaviness of the tails of a distribution. A normal distribution has a kurtosis of 3. Heavy tailed distributions will have kurtosis greater than 3 and light tailed distributions will have kurtosis less than 3.

Source: https://stats.oarc.ucla.edu/stata/output/descriptive-statistics-using-the-summarize-command

10. Different statistical test

- Detail: https://julhas.com/jsedutech/materials/Level-1/Stata-Session-11.pdf

11. Graphics

- Pie chart
  graph pie, over(gender)
  graph pie, over(ageGroup) plabel(_all percent)
  graph pie, over(ageGroup) plabel(_all name)

- Bar chart
  graph bar (mean) numeric_var, over(cat_var)
  graph bar (mean) age, over(gender)

- Histogram
  hist age, freq

- Normal curve with histogram
  histogram age, width(5) freq normal

- Box plot
  graph box age
  graph box age, over(gender)

12. Linear regression

- Example: https://julhas.com/jsedutech/materials/Level-1/Stata-Session-12.pdf

13. Logistics regression

- Example: https://julhas.com/jsedutech/materials/Level-1/Stata-Session-13.pdf