

Tutorial: Exploratory data analysis in Stata

Knowing your data is a key phrase in statistics. Before conducting any sophisticated analyses, first acquaint yourself with the dataset and the variables within the dataset. We go through a sample exploratory data analysis, highlighting key commands that will arise throughout the course.

- **Describe** the variables in the dataset.
 - `describe`
 - `codebook`
- What is the **mean** systolic blood pressure at exam 1 in the study population (`sysbp1`) ?
 - `summarize sysbp1`
- What is the **standard deviation** of `sysbp1`?
- What is the **median** (50th percentile) of `sysbp1`?
 - `summarize sysbp1, detail`
- What is the **range** of `sysbp1`?
 - To use Stata as a calculator, use the command `display`.
- What is the mean systolic blood pressure in males at exam 1?
 - `summarize sysbp1 if sex1==1`
 - `bysort sex1: summarize sysbp1`
- At exam 1, what **percent** of the study population is female?
 - `tabulate sex1`
 - `tabulate sex1, missing`
- How many women have prevalent hypertension at exam 1?
 - `tabulate sex1 prevhyp1`
- **List** the first 50 observations for `sbp1` and `sex1`. Hit the space bar in order to scroll through the observations.
 - `list sysbp1 sex1 in 1/50`

Exploratory data analysis commands you should be familiar with:

`summarize`, `tabulate`, `describe`, `codebook`, `list`