# Supplementary Handout for Dis 3: Sampling; Stata Review

## 1 Sampling

- Recall from discussion 1 that a sample is a subset of data taken from the population. The action of taking this subset of data to construct your sample is called **sampling**.
- Eventually, our goal is to use the sample to draw conclusion about the population (inferential statistics), so it is important that our sample resembles the population.
- Different **sampling plans** are proposed to construct the sample, weighing the benefits of the plan against the costs:
  - **Simple random sampling**: every possible sample entry has equal chance of being selected.
  - **Stratified random sampling**: separate the population into mutually exclusive sets (i.e. strata), and then draw simple random samples from each stratum.
  - **Cluster sampling**: population is first divided into groups, and then one uses simple random sampling to select groups; all observations within the selected groups thus enter the sample.

Exercise. Which sampling plan is used in each of the following examples?

1. Categorize all Econ 310 students based on their class standing (freshman, sophomore, junior, senior, above senior), and then randomly selects 30 students from each class.
2. Categorize all Econ 310 students based on their class standing (freshman, sophomore, junior, senior, above senior), and then randomly select 2 out of the 5 possible groups. The groups corresponding with the class standing selected are chosen as the sample.
3. Number Econ 310 students sequentially from 1 to $N$. Draw 50 non-repeat random positive integers that are less than or equal to $N$. Select the students with the same numbers.

- As you can already see from the exercise, factoring in the specific steps taken when sampling, some sampling plan is expected to construct a sample that more closely resembles the population than the others.

  To formally examine how far the samples are from the population, we look at two types of errors that occur:

  1. **Sampling error**: difference between the sample and the population that exists only because the observations that happen to be included in the sample.
     ⇒ increasing the sample size reduces this error
  2. **Nonsampling errors**: more serious type of error due to samples being selected improperly.
     ⇒ increasing the sample size will NOT reduce this type of error
     Nonsampling errors can be divided into three categories:

(a) **Errors in data acquisition**: the data is recorded wrong (due to incorrect measurement, mistake made during transcription, human errors)

(b) **Nonresponse errors**: responses are not obtained from certain people.

(c) **Selection bias**: some members from the target population cannot possibly be selected to be within the sample.

Exercise. Which type of error arises from the following examples?

1. You sent out a survey to all Econ 310 students via email, but some people quickly archived your email without filling out the survey.

2. You sent out a survey to all Econ 310 students via email, but some freshmen has yet to activate their UW email account, so the survey was not delivered to them.

3. You randomly selected 30 Econ 310 students to have them answer your survey questions. All 30 of them responded, and you did not make any mistake in recording the data. However, your result derived from the sample is still quite different from the parameter in the population.

4. You randomly selected 30 Econ 310 students to have them answer your survey questions. All 30 of them responded, but you messed up the order of items in two columns of the data recorded.

## 2 (Basic) Stata Review

We will briefly go through how you can install Stata in section. Once you have Stata installed, here are some steps that you should take (I'll demonstrate each of them) each time before carrying out your statistical analysis in Stata:

1. Know how the software looks like
2. Tell Stata which folder on your computer you're currently working with (i.e. change your working directory)
3. Import your data using the `use` command, or the `import` command
   - If your dataset ends with `.dta`, the `use` command is appropriate.
     For example, to load the `auto.dta` file, try `use "auto.dta", clear`
   - If your dataset ends with anything but `.dta`, the `import` command is appropriate.
     For example, to load the `auto.xlsx` file, try `import excel "auto.xlsx", firstrow clear`
4. Use a do-file to organize your commands
5. Try using the `help` command and Google to answer your questions

Let's now move on to the Stata handout written by the Econ department to try out some basic commands.