Elementary Statistics Lecture 1

Chong Ma

Department of Statistics
University of South Carolina
chongm@email.sc.edu

Outline

1. Introduction

2. Gathering Data

What is statistics?

- The art and science of learning from data,i.e., the study of a collection, analysis, interpretation and organization of data. The ultimate goal is to translate data into knowledge and understanding the world around us.
- Partly empirical and partly mathematical involving probability theory, measure theory and other related mathematics. Nowadays statistical is more computational.
- Popular statistical softwares: R, SAS, Python, Julia, Minitab . . .
- Broad application: machine learning(Google DeepMind), Biomedical, genetics, econometrics, statistical physics, chemistry, . . .

Questions

Suppose you are assigned a task which is to figure out the situation of American's opinion on the issue that whether abortion should be legalized. How do you do with this?

Some jargons

- **Population**: The total set of subjects in which we are interested. Could be “All people living in SC” or “every atom composing a crystal”
- **Sample**: The subset of the population for whom we have data, often random selected.
- **Descriptive Statistics**: methods for summarizing the collected data, usually consisting of graphs and numbers.
- **Inference Statistics**: methods for making decisions or predictions on the population, based on data obtained from a sample of that population.

Continue...

- **Parameter**: A numerical summary of the population.
- **Statistic**: A numerical summary of a sample taken from the population.
- **Random Sampling**: Make the sample representative of the population, i.e., each subject in the population has the same chance of being included in that sample.
- **Margin of Error**: a measure of the expected variability from one random sample to another random sample.

Example 1

Sleep disorders among college students An article in *journal of American College Health* reports that, in a survey of 1845 college students from a large, southeastern public university, 27% were at risk for at least one sleep disorder, with a margin of error 2%.

- **Population**: All college students in that University.

Example 1

Sleep disorders among college students An article in *journal of American College Health* reports that, in a survey of 1845 college students from a large, southeastern public university, 27% were at risk for at least one sleep disorder, with a margin of error 2%.

- **Population**: All college students in that University.
- **Sample**: 1845 college students in the survey.

Example 1

Sleep disorders among college students An article in *journal of American College Health* reports that, in a survey of 1845 college students from a large, southeastern public university, 27% were at risk for at least one sleep disorder, with a margin of error 2%.

- **Population**: All college students in that University.
- **Sample**: 1845 college students in the survey.
- **Sample Statistics**: 27%

Example 1

Sleep disorders among college students An article in *journal of American College Health* reports that, in a survey of 1845 college students from a large, southeastern public university, 27% were at risk for at least one sleep disorder, with a margin of error 2%.

- **Population**: All college students in that University.
- **Sample**: 1845 college students in the survey.
- **Sample Statistics**: 27%
- **MOE**: $2\% \approx \frac{1}{\sqrt{n}} = \frac{1}{1845}$

Example 2

At what age did women marry A historian wants to estimate the average age at marriage of women in New England in the early 19th century. Within her state archives she finds marriage records for the years 1800-1820, which she treats as a sample of all marriage records from the early 19th century. The average age of the women in the records is 24.1 years. Using the appropriate statistical method, she estimates that the average age of brides in early 19th century New England was between 23.5 and 24.7.

- **Population**: Married women in New England in early 19th century.

Example 2

At what age did women marry A historian wants to estimate the average age at marriage of women in New England in the early 19th century. Within her state archives she finds marriage records for the years 1800-1820, which she treats as a sample of all marriage records from the early 19th century. The average age of the women in the records is 24.1 years. Using the appropriate statistical method, she estimates that the average age of brides in early 19th century New England was between 23.5 and 24.7.

- **Population**: Married women in New England in early 19th century.
- **Sample**: Women in the records for the years 1800-1820.

Example 2

At what age did women marry A historian wants to estimate the average age at marriage of women in New England in the early 19th century. Within her state archives she finds marriage records for the years 1800-1820, which she treats as a sample of all marriage records from the early 19th century. The average age of the women in the records is 24.1 years. Using the appropriate statistical method, she estimates that the average age of brides in early 19th century New England was between 23.5 and 24.7.

- **Population**: Married women in New England in early 19th century.
- **Sample**: Women in the records for the years 1800-1820.
- **Descriptive summary**: The average age of the women in the records is 24.1 years.

Example 2

At what age did women marry A historian wants to estimate the average age at marriage of women in New England in the early 19th century. Within her state archives she finds marriage records for the years 1800-1820, which she treats as a sample of all marriage records from the early 19th century. The average age of the women in the records is 24.1 years. Using the appropriate statistical method, she estimates that the average age of brides in early 19th century New England was between 23.5 and 24.7.

- **Population**: Married women in New England in early 19th century.
- **Sample**: Women in the records for the years 1800-1820.
- **Descriptive summary**: The average age of the women in the records is 24.1 years.
- **Inference**: She estimates that the average age of brides in early 19th century was between 23.5 and 24.7.

Outline

1. Introduction

2. Gathering Data

Experimental and Observational Studies

1. Types of Studies
   - Experimental Study: Assign subjects to certain experimental conditions and then observing outcomes on the response variables.
   - Observation Study: Observe values of the response variable and explanatory variable for the sampled subjects.
2. Comparison
   - Experimental study has advantages of establishing cause and effect than observation study, by ruling out lurking variables as much as possible.

Remark

- Response variable: the outcome of interests
- Explanatory variable: related to response variable in the study.
- lurking variable: not observed in the study that influences the association between the response and explanatory variables due to its own association with each of those variables.

Sampling methods(Observational Study)

I **Probability Sampling**

1. Simple Random Sample(SRS)
2. Stratified Sampling
3. Cluster Sampling
4. Systematic Sampling
5. Multistage Sampling(some of methods above combined in a stage)

II **Non-probability Sampling(poor way)**

1. Convenience samples
2. Volunteer samples

Sampling methods(Observational Study)

Simple Random Sample(SRS): A SRS of n subjects from a population is on in which each possible sample of that size has the same chance of being selected.

Definition

The sampling frame is the list of subjects in the population from which the sample is taken.

Exercise 1

Conduct a SRS of 6 students from our class of 72 students.

Sampling methods(Observational Study)

Stratified Sampling: partitions the population into groups based on a factor that may influence the variable is being measured.

- partition the population into groups
- obtain a SRS from each group (stratum)
- collect data on the random sampling subjects from each group

	Example 1	Example 2
Population	All people in SC	All STAT 201 students
Groups(Strata)	46 counties in SC	46 sections in USC
Obtain a SRS	20 people from each of the 46 counties	4 students from each of the 46 sections
Sample	$20 \times 46 = 920$	$4 \times 46 = 184$

Table 1: Examples of Stratified Samples

Sampling methods(Observational Study)

Cluster Sampling: the clusters are microcosms, rather than subsections of the population.

- divide the population into groups (clusters)
- obtain a SRS of so many clusters from all possible clusters
- collect data on every sampling subject in each of the randomly selected clusters.

	Example 1	Example 2
Population	All people in SC	All STAT 201 students
Groups(Clusters)	46 counties in SC	46 sections in USC
Obtain a SRS	3 counties from the 46 possible counties	4 sections from the 46 possible sections
Sample	every person in the 3 selected counties	every students in the 4 selected sections

Table 2: Examples of Cluster Samples

Sampling methods(Observational Study)

Types of Bias

- **Sampling bias** occurs from non-random samples or having undercoverage.
- **Nonresponse bias** occurs when some sampled subjects cannot be reached or refuse to participate or fail to answer some questions.
- **Response bias** occurs when the subject gives an incorrect responses (perhaps lying) or the way the interviewer asks the questions is confusing or misleading.

Good ways to Experiment

- Set a control comparison group and a treatment group.
- Blindingly and randomly assign experimental units to the control and treatment group.

Role of randomization

- To eliminate bias that may result if you assign the subjects
- To balance the groups on variables that you know affect the response
- To balance the groups on lurking variables that may be unknown to you

Example 3

Antidepressants for Quitting Smoking To investigate wheter antidepressants help people quit smoking, one study used 429 men and women who were 18 or older and had smoked 1 cigarettes or more per day for the previous year. They were randomly assigned to one of two groups: One group took 300 mg daily of an antidepressant that has the brand name bupropion. The other group did not take an antidepressant. At the end of a year, the study observed whether each subject had successfully abstained from smoking or had relapsed.

- **Response Variable**: Whether the subject abstains from smoking for one year (yes or no)

Example 3

Antidepressants for Quitting Smoking To investigate wheter antidepressants help people quit smoking, one study used 429 men and women who were 18 or older and had smoked 1 cigarettes or more per day for the previous year. They were randomly assigned to one of two groups: One group took 300 mg daily of an antidepressant that has the brand name bupropion. The other group did not take an antidepressant. At the end of a year, the study observed whether each subject had successfully abstained from smoking or had relapsed.

- **Response Variable**: Whether the subject abstains from smoking for one year (yes or no)
- **Explanatory Variable**: Whether the subject received bupropion (yes or no)

Example 3

Antidepressants for Quitting Smoking To investigate wheter antidepressants help people quit smoking, one study used 429 men and women who were 18 or older and had smoked 1 cigarettes or more per day for the previous year. They were randomly assigned to one of two groups: One group took 300 mg daily of an antidepressant that has the brand name bupropion. The other group did not take an antidepressant. At the end of a year, the study observed whether each subject had successfully abstained from smoking or had relapsed.

- **Response Variable**: Whether the subject abstains from smoking for one year (yes or no)
- **Explanatory Variable**: Whether the subject received bupropion (yes or no)
- **Treatment**: buropion, no buropion

Example 3

Antidepressants for Quitting Smoking To investigate wheter antidepressants help people quit smoking, one study used 429 men and women who were 18 or older and had smoked 1 cigarettes or more per day for the previous year. They were randomly assigned to one of two groups: One group took 300 mg daily of an antidepressant that has the brand name bupropion. The other group did not take an antidepressant. At the end of a year, the study observed whether each subject had successfully abstained from smoking or had relapsed.

- **Response Variable**: Whether the subject abstains from smoking for one year (yes or no)
- **Explanatory Variable**: Whether the subject received bupropion (yes or no)
- **Treatment**: buropion, no buropion
- **Experimental units**: The 429 volunteers who are the study subjects

Example 4

In 1950 in London, England, medical statisticians Austin Bradford hill and Richard Doll conducted one of the first studies linking smoking and lung cancer. In 20 hospitals, they matched 709 patients admitted with lung cancer in the preceding year with 709 noncancer patients at the same hospital of the same gender and within the same five-year grouping on age. All patients were queried about their smoking behavior. A smoker was defined as a person who had smoked at least one cigarette a day for at least a year.

	Lung Cancer	
Smoker	**Yes(case)**	**No(Control)**
Yes	688	650
No	21	59
Total	**709**	**709**

Table 3: Results of retrospective study of smoking and lung cancer