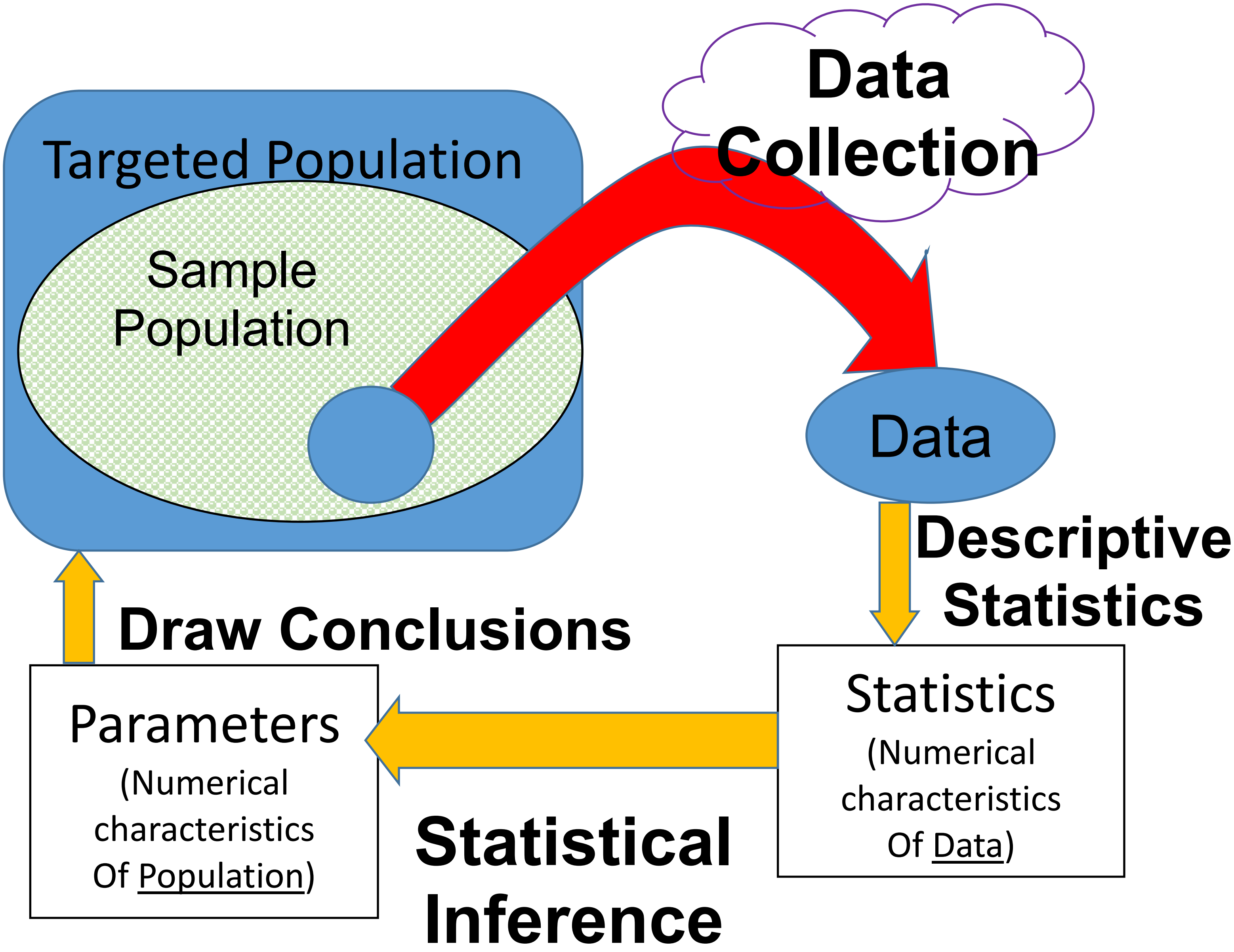
MATH 7343 Applied Statistics

Notes

Overview of the statistical data analysis process:

1. Define the problem in the statistics framework.
2. Collect data.
3. Analyze data, interpret and communicate the results.



Definitions:

Population: A collection of items or individuals in which we are interested.

Sample (Data): A subset of the population that is selected or available for taking measurements.

Parameter: A numerical description of some characteristic of the population.

Statistic: A numerical description of some characteristic of the sample.

Target population: The population about which we ultimately wish to make inference.

Sampled population: The population from which the sample is taken.

Example 1. *Political polling.*

A 2004 Massachusetts poll was conducted to determine the public opinion about the constitution amendment to ban same-sex marriage. From the public phone directory of the Massachusetts area, 376 names were selected and those people were called and asked about whether they support the amendment or not. Of those, 329 person actually responded to the survey, of which 173 persons supported the amendment.

Target population: All Massachusetts residents.

Sampled population: People who had a publicly listed phone number in Massachusetts and were able and willing to take the survey.

Sample: The 329 persons who responded.

Parameter: The proportion of Massachusetts residents that support the marriage amendment.

Statistic: 173 out of 329 support the marriage amendment.

Discussion: Is the sample representative of the population?

Example 2: R usage by graduate students.

Suppose I am interested in how many Northeastern University graduate students used the R software before. I ask everyone in class to tell me if he/she has used R before.

Target population:

Sampled population:

Sample:

Parameter:

Statistics:

Discussion: Is this sample representative of the targeted population? What is a more appropriate targeted population?

Example 3. ***Election polling.***

A Newsweek poll was conducted from September 2-3, 2004 to predict the result of 2004 Presidential election. A nationwide phone survey of 1008 registered voters was asked their preference in the election. Of those, 524 persons planned to vote for G.W. Bush and 413 planned to vote for J. Kerry.

Target population:

Sampled population:

Sample:

Parameter:

Statistic:

Discussion: Is the sample representative of the population? How to improve the poll?

There was an intense discussion whether the double-digit lead by Bush after Republican convention is real or not. Particularly, the Newsweek poll sample has 374 Republicans, 303 Democrats and 300 Independent. (On election date later, a CNN poll shows that party registration among voters are 37% Republican, 37% Democrat and 26% Independent.)

The mismatch between the party affiliations among the sample versus among the voter registration was often raised as an issue. For example, http://www.emory.edu/news/Releases/gallup1094590402.html

Checking representativeness of the sample: parameters?   Statistics?

How to correct? Weighting by party affiliations is one way. But that may not work well in practice. See the following link. https://www.theguardian.com/commentisfree/2012/sep/17/weighting-polls-party-identification

Major Techniques of Inference
(Just an illustration here. We will learn those in detail later in the course.)

Point Estimate: provides a best guess of the value of a parameter.

Interval Estimate: provides an interval of values in which we fell with some certainty that the parameter lies.

Hypothesis Test: a decision making procedure to decide if a statement about the parameter (called a null hypothesis) is true or false.

Significance Test: a measure of the plausibility of a null hypothesis in light of the data.

Example 1. *Political polling.* (Cont'd)

Let us consider the proportion of the marriage amendment supporters.

Point Estimate: $\hat{p} = x/n = 173/329 \quad =$

Interval Estimate: $\hat{p} \pm Z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n} = \qquad \pm 1.96\sqrt{\frac{(\quad)(\quad)}{\quad}} = (\qquad , \qquad)$

Hypothesis Test: Does the marriage amendment has the majority support in Massachusetts?

$H_0: p \leq 0.5$

Reject H$_0$ if $\hat{p} \geq p_0 + Z_{\frac{\alpha}{2}}\sqrt{\frac{p_0(1-p_0)}{n}} = 0.5 + 1.64\sqrt{\frac{0.5(1-0.5)}{329}} = 0.545$

Answer: Accept H0. There is not enough evidence that the majority Massachusetts residents support the marriage amendment.

Significance Test:
p-value$=1 - \Phi\left(\frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/n}}\right) = 1 - \Phi(\frac{0.526-0.5}{\sqrt{0.5(1-0.5)/329}}) = 1 - \Phi(0.94)$ =0.17

More concepts.

Random sampling: A method for selecting a sample from a population such that every member of the population has an equal chance of being selected.

A sample is a random sample form a population if
(1) Each data value is selected in a manner totally unrelated to the other data values in the sample (independence).
(2) Each data value is selected from a set of values that is the same as those in the population (identically distributed).

Example 1. *Political polling.* (cont'd)

To ensure the sample is a random sample, a table of random digit is used. (Or more likely, a computer algorithm generating random numbers is used.) Could see chapter 22 if you are interested in more sampling theory.

An example of Non-random sample could be that the first 376 names in the phone directory being chosen. We would then be stuck with all persons with the last name starts with an "A". This could leads to many people from the same racial group or even the same family and rendering any conclusion from the sample useless for the general population.

Association is NOT Causation

Association: There is a relationship between two random variables X and Y. For example, “owning an iPhone in 2016” is associated with “having higher income” according to this paper.

==

Researchers find that owning an iPhone or iPad is the number-one way to guess if you’re rich or not

Kif Leswing
Jul. 8, 2018, 11:00 AM AP

In the United States, if you have an Apple iPhone or iPad, it's a strong sign that you make a lot of money.

……

Causation: Changes in one variable X directly causes changes in another variable Y.
Would buying an iPhone in 2016 cause you to have higher income?

You may also read the following paper “Association, correlation and causation” (https://www.nature.com/articles/nmeth.3587) in Nature Methods.

Do we need a causation result? Is associative result ok?
(Answer: depends on the study objective: marketing research? Scientific study? ….)

Can we use statistical studies to infer if X is associated Y or X causes Y?
Association can be judged statistically if the study is properly designed (random sampling).
Causation needs stronger controlled design.

Controlled vs. Observational in comparison studies.

In a controlled experiment the allocation of “treatment” to subjects is done at random. Randomization eliminates biases and *enables the inference of causation.*

In an observational study the “treatment” is a characteristic of the subject which we merely observe. It is usually not possible to eliminate biases and we can *only infer association*.

Example 4. Controlled vs. Observational

Question: What is the effect of forest fires on the density of oak seedlings in a national forest?

Experiment: Count the number of oak seedlings in a 10 square meters plot in the forest.

Controlled experiment

1. Select sites to study.
2. Choose half of sites at random and burn them.
3. Return later to count the seedlings at the sites.

Observational Study

1. Select the sites to study such that some are burned and some are not.
2. Count the seedlings at the selected sites.

Example 5. Video display terminals and miscarriage. (Observational study)

Question: Does exposure to video display terminals increase risk of miscarriage?

Target population: All pregnant women in the U.S. over next several years.

Sample: 1583 pregnant women who attended one of the three Boston area hospitals.

Statistics: Difference in miscarriage rates between women exposed to VDTs and women not exposed.

Conclusion:

Example 6. Polio Vaccine effectiveness. (Controlled experiment)

In 1950s, a Polio vaccine was tested on all elementary school children whose parent gave consent.

Target Population: All children in the US over next several years.

Sample: The 400,000 plus children in the trial.

Statistics:
Vaccine group: 33 of 200,745 developed paralysis. Rate 16/100000
Placebo group: 115 of 201,229 developed paralysis. Rate 57/100000

Conclusion:

Experimental Unit: The largest unit to which a single treatment is applied.
Sampling Unit: The largest unit on which measurements are made.
(Variance among treatments can only be reflected through variance among different experimental units. If sampling units are different from experimental units, measurement variance and treatment variance need careful distinction.)

Examples:

1. The Professor pops two bags of regular and two bags of gourmet popcorn. One bag of each type has been stored at room temperature, while the other has been frozen. He passes the popped corn around the class, and each student takes a handful and consumes them.

Experimental unit: a bag of popcorn.
Sampling unit: a handful of popcorn.

Is it possible to distinguish the treatment effect, with large number of sampling units in each experimental units?

2. A farmer puts five experimental fertilizers (including one control) on ten different plots in a field. Each fertilizer is placed on two plots. He then plants corn on each of the plots.

(a) At the end of the growing season, he selects ten plants from each plot and weighs them individually.
experimental unit: a plot of corn.
sampling unit: an individual corn plant.

(b) At the end of the growing season, he harvests and weighs all the corn on each plot.
experimental unit: a plot of corn.
sampling unit: a plot of corn.

Type of Data:

Quantitative data are measurements or counts that have meaningful numerical values.

Qualitative data are attributes such as gender, occupation or the responses to a yes/no question that do not have inherent numerical values. Usually represented as nominal, ordinal or ranked data.

Continuous data are measurements that could in principle be made arbitrarily precise. For example, weight or temperature.

Discrete data belong to distinct classes and are usually counts or qualitative.