MATH 7343 Applied Statistics

Prof. (Aidong) Adam Ding

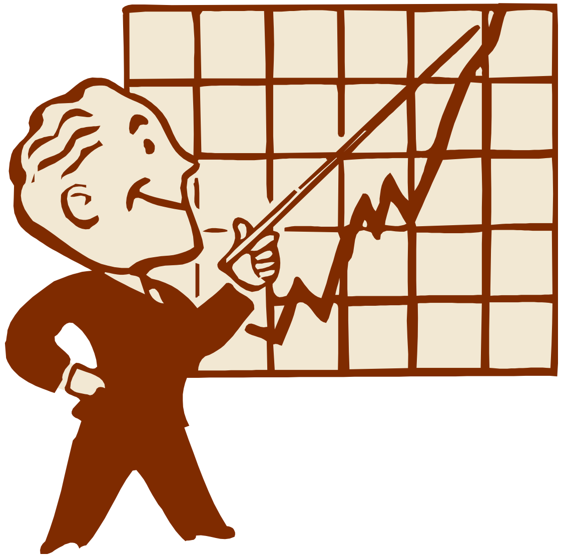
Northeastern University

First day:

Administrative items, Statistical concepts

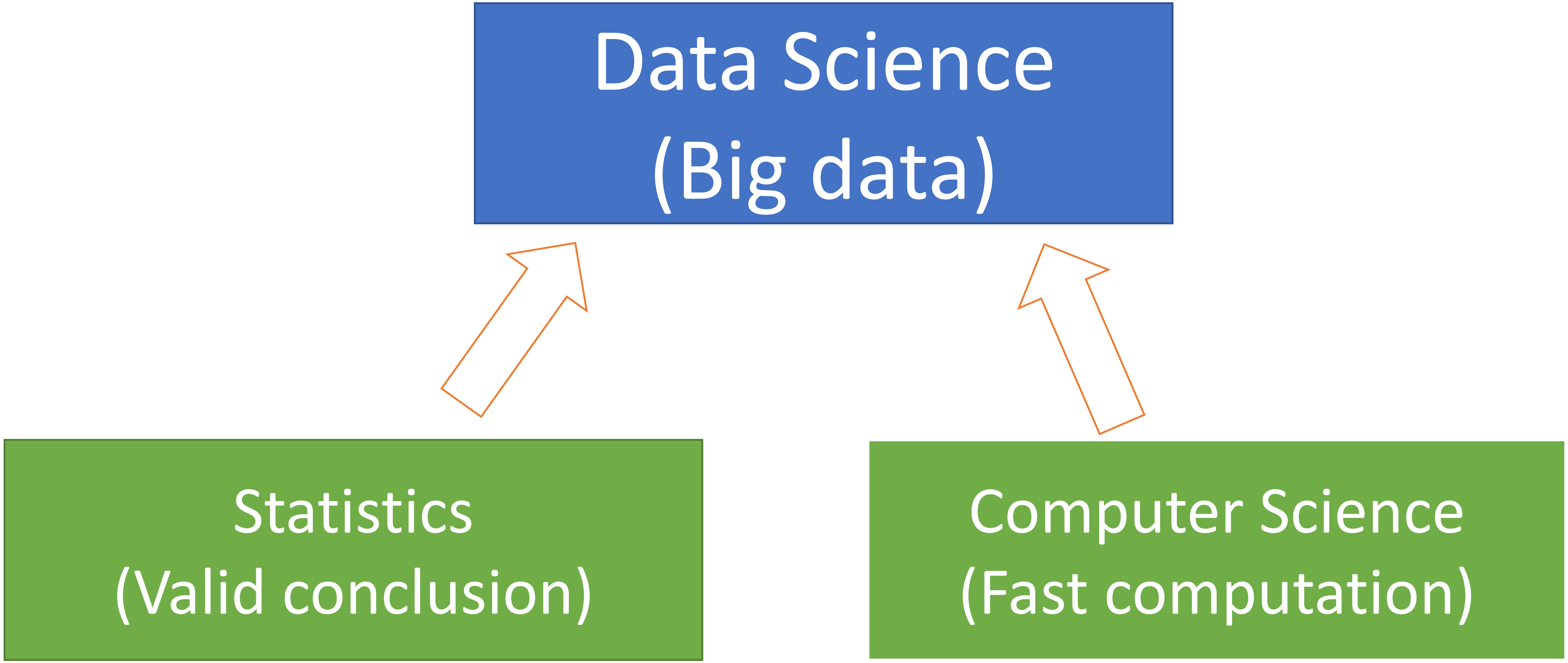
- First, go over syllabus. Learn to install R on your own.
- What is statistics?
- Basic statistical terms and concepts

STATISTICS



A form of Mathematics concerned with methods for

Collecting, Organizing, Presenting, and Analyzing Data, as well as *drawing valid conclusions* and *making reasonable decisions* on the *basis of such analyses.*



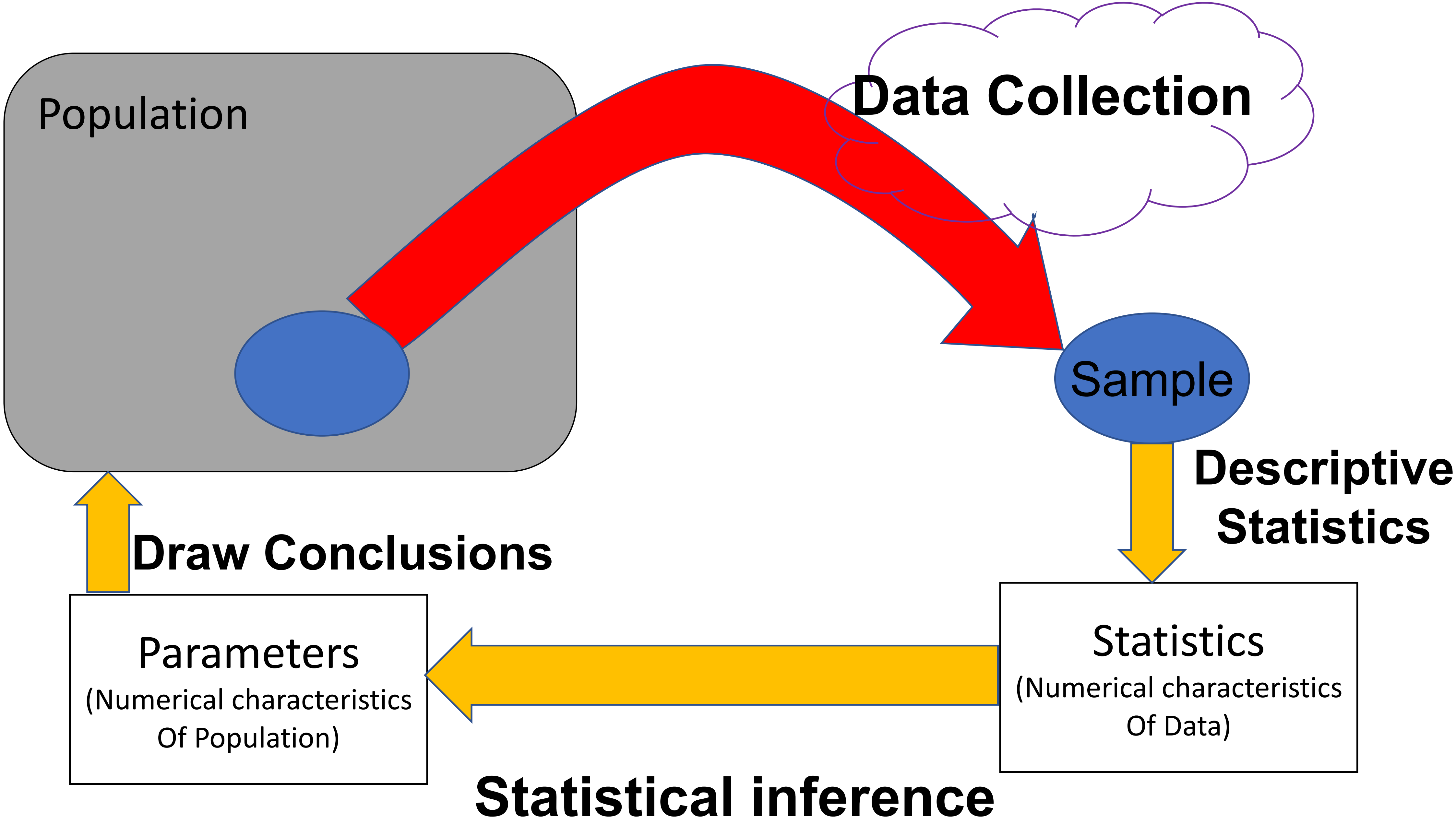
Data Science
(Big data)

Statistics
(Valid conclusion)

Computer Science
(Fast computation)

Overall process of statistical analysis.

1. Define problem
2. Collect data
3. Describe and presentation
4. Draw conclusion


Population
Data Collection
Sample
Descriptive Statistics
Draw Conclusions
Parameters
(Numerical characteristics Of Population)
Statistics
(Numerical characteristics Of Data)
Statistical inference

Basic statistical terms (concepts).

- 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.

Overall process of statistical analysis.

1. Define problem
2. Collect data

Sampling theory; Experimental Design.

3. Describe and presentation

Descriptive statistics

4. Draw conclusion

Inferential statistics

What we want to learn from the data

Information:

Data summary (descriptive statistics)

versus

Knowledge:

Statistical Inference

Overall process of statistical analysis.

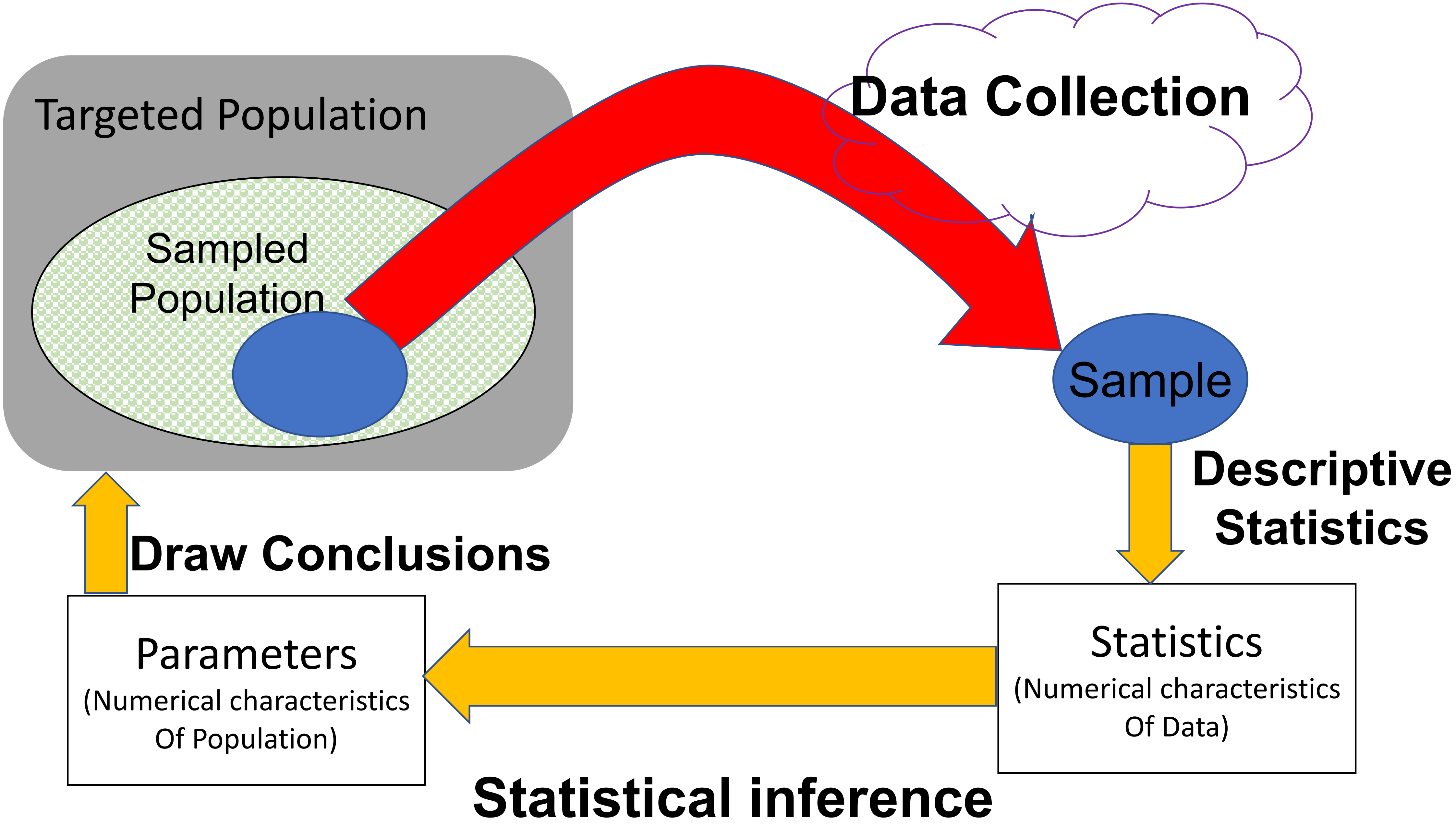
1. Define problem
2. Collect data
3. Describe and presentation
4. Draw conclusion

Majority of this course focused on *STEP 4 inferential statistics*.

However, statistical consideration are needed in *every step* of this process.

A closer look at the population

- Target Population: The population about which we ultimately wish to make inference.
- Sampled population: The population from which the sample is taken.


Targeted Population
Sampled Population
Data Collection
Sample
Descriptive Statistics
Statistics
(Numerical characteristics Of Data)
Statistical inference
Parameters
(Numerical characteristics Of Population)
Draw Conclusions

Can we make all statistical inferences objective?

- Is the sample representative of the population?

(Make it objective using sample theory)

- Is the sampled population representative of the target population?

(Has to use some subjective judgement)

Concepts 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.

- Identify population/sample/parameter/statistic/

Concepts 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.

Concepts 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.

- **Discussion:**

  Is the sample representative of the population?

Concepts Example 2. *R usage by graduate students.*

I am interested in how many Northeastern University graduate students used the R software before. Everyone here please answer the poll if he/she has used R before.

- Target population:
- Sampled population:
- Sample:
- Parameter:
- Statistic: 5 out of 25 students have used

Concepts Example 2. *R usage by graduate students.*

I am interested in how many Northeastern University graduate students used the R software before. Everyone here please answer the poll if he/she has used R before.

- **Discussion:**

  Is the sample representative of the population?

Concepts 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:

Concepts Example 3. *Election polling.*

Discussion: Is the sample representative of the population?

Can we 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

Concepts Example 3. *Election polling.*

Discussion: Is the sample representative of the population?

- We can statistically check if the proportions of party registration are same in the sample and population.
- The Newsweek poll sample has 374 Republicans, 303 Democrats and 300 Independent. Versus party registration among all voters of 37% Republican, 37% Democrat and 26% Independent.
- Parameter: Rep/Dem/Ind proportions in population (37%, 37%, 26%)
- Statistics: Rep/Dem/Ind proportions in sample (374, 303, 300)/977

Note this is a different question from the original research problem.

- Sample NOT representative.

Concepts Example 3. *Election polling.*

Discussion: Is the sample representative of the population?

- The Newsweek poll sample has 374 Republicans, 303 Democrats and 300 Independent. Versus party registration among all voters of 37% Republican, 37% Democrat and 26% Independent.
- Sample NOT representative.
- How to correct? Weighted estimates?

- Weighting by party affiliations may not work well in practice. See https://www.theguardian.com/commentisfree/2012/sep/17/weighting-polls-party-identification

Major Techniques of Inference

- 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.

Major Techniques of Inference

- Next two slides shows how these techniques lead to statistical *inferences* (drawing conclusions)

- Do not worry about the actual mathematical details, which we will go over during the semester. These are just illustrations on how the mathematical formulas can be used to draw valid conclusions.

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

- Parameter: The proportion of Massachusetts residents that support the marriage amendment.
- Statistic: 173 out of 329 support the marriage amendment.

- <u>Point Estimate:</u> $\hat{p} = x/n = 173/329 =$

- <u>Interval Estimate:</u> $\hat{p} \pm Z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n} = \quad \pm Z_{\alpha/2}\sqrt{\quad (1- \quad)/n}$

  $= (\quad , \quad)$

Concepts Example 1. *Political polling. (Cont’d)*

- Statistic: 173 out of 329 support the marriage amendment.
- Point Estimate: $\hat{p} = 52.6\%$
- 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=$1 - \Phi\left(\frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/n}}\right)$= $1 - \Phi(\frac{0.526-0.5}{\sqrt{0.5(1-0.5)/329}})$

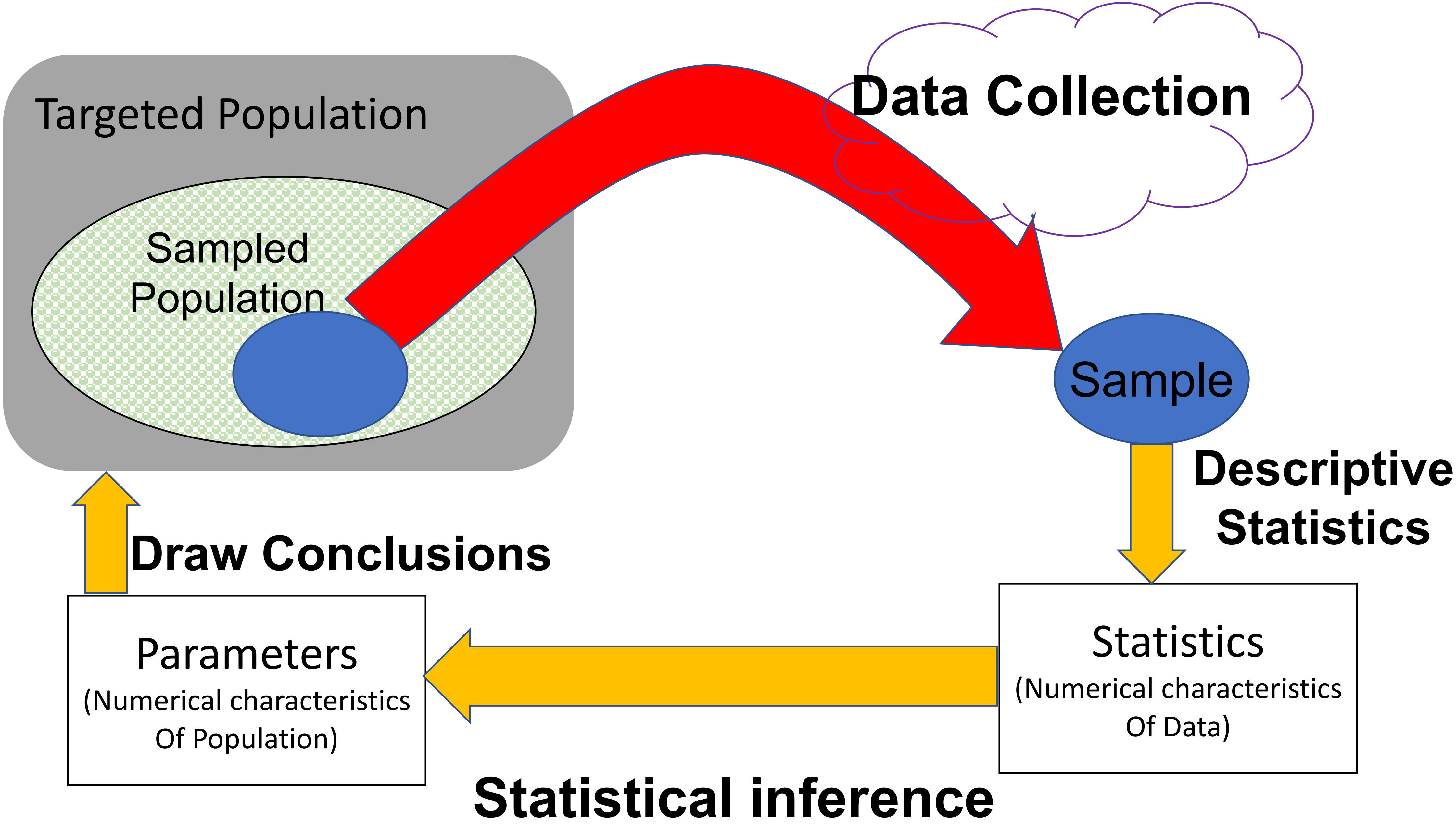
=0.17

Summary of first lecture

- Basic concepts of statistics:

Most of the course focus on the statistical inferences: draw conclusions about the parameter (of population) using the statistics (of sample).

Inferential statistics is only a part of the process of statistical analysis, the other parts also involves statistical considerations (may even need subjective judgement). It is critical that all phrases of the statistical analysis follow sound procedures.


Targeted Population
Sampled Population
Data Collection
Sample
Descriptive Statistics
Statistics
(Numerical characteristics Of Data)
Statistical inference
Parameters
(Numerical characteristics Of Population)
Draw Conclusions

Within a week

- Install R on your PC

- Fill out the google form for project collaboration (answer the background questions, which I will use to assign people to group. If you have classmates whom you want to be in the same group, please write the names down. Otherwise, leave that question blank.)