Introduction to ST372:

The discipline of Statistics is concerned with

1. collecting data,
2. describing data (which may have been collected by someone else),
3. drawing formal inferences from data (where appropriate),
4. while quantifying and acknowledging natural variability.

Why we want to use data—and what do we mean by data?

- We have a question we want to answer, and then apply what we learn in a new situation, or use what we learn to make a decision.
- Data generally refers to the same variable collected for many “things,” as opposed to information—which is, say, one person’s phone number.

Population, sample, and census

Population: the entire collection of “things” we are interested in

Enumerative studies: Have well-defined populations.
- College students
- Citizens of Iraq
- The students in my ST372 class

Analytic studies: Have conceptual or hypothetical populations.
- The conceptual population of hair shampooed with this new product.
- The conceptual populations of balsa wood bridges made with each of the following 6 specifications.
- The conceptual population of bearings manufactured with this new process
- The conceptual population of bearings manufactured with the old process

Sample: a subset of the population for which we have data

Census: a study in which the sample is equal to the population—like when I get all of you to sign the attendance sheet and tell me what your major is—easy when the population is small
**Where people get data:**

1. **Experiments (some are more useful than others)**
   - When you want to know the effect of different explanatory variables on a response variable.
     - I have drug A, drug B, and no drug at all. Which treatment makes people with depression feel better?
     - I want to manufacture bearings with a new process. Will it result in less road noise than the old process?
   - *A well-designed* experiment is the gold standard for drawing inferences about cause-effect relationships.
   - In this class, you will have a group project in which you will be designing an experiment, implement your design, and analyze the data, all using formal statistical methods.

2. **Observational study**
   - When you want to know something about a population but experimentation would be inappropriate,
   - When the question is not about a response-to-change (opinions, physical characteristics, health information)
   - *Probability samples* are the best—well designed for the intended purpose.
   - A sample that isn’t a probability sample (sometimes ok, sometimes bad, section 3.3)

3. **Anecdotal evidence**
   Based on haphazardly selected individual cases, which often come to our attention because they are striking in some way. They may or may not be representative of a larger group of cases.

4. **Available data**
   Data that were produced in the past for some other purpose but that may help answer a present question.
### Exploratory data analysis/descriptive statistics vs. formal statistical inference

<table>
<thead>
<tr>
<th>Why would we use it?</th>
<th>Exploratory data analysis/Descriptive statistics</th>
<th>Formal statistical inference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>• Dataset was not collected using a well-designed experiment or sample.</td>
<td>• The question is important.</td>
</tr>
<tr>
<td></td>
<td>• Data was not collected specifically for the purpose of answering the question at hand.</td>
<td>• The data were collected specifically for the purpose of answering one or more questions about the population of interest.</td>
</tr>
<tr>
<td></td>
<td>• Dataset is too small for formal statistical inference about a population.</td>
<td>• The data were collected using a well-designed experiment and/or sample.</td>
</tr>
<tr>
<td></td>
<td>• Dataset does not represent a random sample of the population of interest.</td>
<td>• Allows us to give a measure of confidence in our inference.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Allows us to quantify uncertainty</td>
</tr>
<tr>
<td>What does it look like?</td>
<td>• Graphical depictions of the dataset.</td>
<td>• Graphical depictions of dataset (always).</td>
</tr>
<tr>
<td></td>
<td>• Statements about the dataset.</td>
<td>• Statistical tests, designed to answer a specific question, based on assumptions about the probability distribution of the population. (The probability distribution is a model of the population.)</td>
</tr>
<tr>
<td></td>
<td>• Statements about the sample.</td>
<td>• Inference (conclusions) about the population of interest.</td>
</tr>
<tr>
<td></td>
<td>• No inferences (conclusions) about a broader population, because sample does not represent a random sample from broader population.</td>
<td>• Answer to the important question about the population of interest.</td>
</tr>
</tbody>
</table>
HW: Remembering what you should know from prerequisite—use the text, either 7th or 8th edition

1. Fill in the Table of Analogies as best you can (some of it requires mind reading, but some not)—we will go over this next time.
2. Get your TI83, 84, or whatever—you will be using it immediately!
3. Download JMP onto your personal computer—you will be using it after next class, so get this taken care of now.