Lecture 2 (CR.1)

Statistics (at this level) is NOT math!
But there is a lot of math in it.

It is extremely ambiguous. How wide is this curve?
It is more like a language.

“At the 95% confidence level, the observed confidence interval covers the true population regression fit at a given x."

Two Types of Statistics:

<table>
<thead>
<tr>
<th>Descriptive</th>
<th>Inferential</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
<td>To infer something about</td>
</tr>
<tr>
<td>median</td>
<td>a population = Technical term</td>
</tr>
<tr>
<td>mode</td>
<td>from</td>
</tr>
<tr>
<td>range</td>
<td>a Sample = Technical term</td>
</tr>
<tr>
<td>histogram</td>
<td>to which we have incomplete access.</td>
</tr>
<tr>
<td>scatterplot</td>
<td>A subset to which we have complete access.</td>
</tr>
</tbody>
</table>

Note: Everytime you collect data, you are really taking a sample from a population. So, in practice, you take a sample, describe/summarize it using methods of Descriptive Statistics, and then use Inferential Statistics to say something about the population from which the sample was collected.
Both sample & pop are described in terms of **variables** (e.g. length, mass, ...). There are different types of vars, because each type requires a different methodology for analysis.

1) Quantitative
   a) Continuous \( x \in \mathbb{R} \)
      - \( x = \) time it takes to complete a computer code.
   b) Discrete \( x \in \text{Integers} \)
      - \( x = \) # of defective elements in a computer. \( x \in \{0, 1, 2, \ldots\} \)
      - \( x = \) # of Macs in a class of 100 students. \( x \in \{0, 1, \ldots, 100\} \)

2) Qualitative (or Categorical)
   - \( x = \) computer type in a class. \( x \in \{\text{Mac, Dell, HP}\} \)
   - \( x = \) state of a coin. \( x \in \{\text{Heads, Tails}\} \)
   - \( x = \) letter grades in a class of 120 students. \( x \in \{A, B, C, D, F\} \)

Here is something very important:

**Random Variable** This is a very important concept in statistics. All we need to know about it is that it is a variable (e.g. length, time, type, ...) that changes values every time we observe/measure it in a random sample. So, the word "random" in random var. refers to the sample being random. And because we will assume all samples are random, we often drop the word "random." Things referring to the population (e.g. mean of all \( x \)'s in a population) are NOT r.v.'s.
Data (i.e. sample) on these r.v.'s may look like this:

<table>
<thead>
<tr>
<th>Case</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$x_4$</th>
<th>$x_5$</th>
<th>$x_6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Short</td>
<td>3.14</td>
<td>A</td>
<td>B</td>
<td>Mac</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>medium</td>
<td>2.79</td>
<td>C</td>
<td>B</td>
<td>HP</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>tall</td>
<td>...</td>
<td>B</td>
<td>G</td>
<td>Dell</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>tall</td>
<td>...</td>
<td>C</td>
<td>G</td>
<td>HP</td>
<td>2</td>
</tr>
</tbody>
</table>

This categorization may seem easy, but it's not. Look:

Here are some ambiguities:
- Is $x_2 \times 10,000$ discrete?!

**Answer:** It depends.

- Suppose you observe $x_2$ 100 times, but get:

  1.3, ..., 2.2, ..., 1.7, ..., 0.5, ...

  25 times "" "" ""

Then, it's best to treat $x_2$ as discrete, with 4 levels!

But if we get 100 distinct/different values, then treat it as cont.

What's the cutoff/boundary between discrete and cont?

**Answer:** It depends on, e.g. the total sample size. And/or what you want to do with the data. You will gain some experience in class.

Like I said, it's complicated!
One place where the distinction matters is in histograms.

For Categ/Qual. r.v.'s hists are easy to make:

Just count the # of cases for each level of the variable.

E.g. $x =$ "favorite fruit type" Data on, say, 7 people.

$x \in \{\text{Orange, Apple, Banana, Orange, Kiwi, Orange, Apple}\}$

For Continuous r.v.

Divide up the $x$-axis into some number of intervals/bins, and count how many cases fall in each bin/interval.

E.g. Data: $x = 1.05, 1.25, 1.41, 1.48, 1.75$

In R: see lab1. hist$(x, \text{breaks=...})$

\[ \text{width = 0.01} \]

\[ \text{binsize = 0.1} \]

\[ \text{binsize = 2} \]

\[ \Rightarrow \text{bin too small (useless)} \]

\[ \Rightarrow \text{bin too large (useless)} \]

\[ \text{In lab you learn how to turn the knob to reveal hidden patterns in data (e.g. the existence of 2 groups.)} \]

\[ \Rightarrow \text{shape is important} \]

For discrete r.v.'s: hists can be made with or without bins.

Really, really, powerful data analysis too!
Come-up with 2 examples for each of the three types of variables (continuous, discrete, categorical). As discussed in this lecture, the type of a variable cannot be determined without the actual data, i.e., the type depends on the specifics of data. Here, however, ignore that complexity, and base your answer on theoretical considerations (i.e., based on what you know about that variable).

Construct a data set with the following specifications. Any source is allowed: web, books, papers, your own work, etc. However, the data cannot be made-up! It must pertain to a real problem. We will apply every technique we learn to this data set. Put thought and effort into it, because in the past I have been able to help students to get a journal publication based on their work.

Specifications:
1) Number of cases: 30, or more
2) Two categorical or discrete variables. One of them must have between 2 and 6 levels, and the other must have between 3 and 6 levels. See part b) for a requirement on the histograms.
3) Two continuous variables.
4) The four variables must relate to a common problem, not four unrelated problems.

a) Print the data in the following format, and turn it in.
column1 = var1, column2 = var2, etc..

b) Plot histograms for each of the four variables. By R.
For the continuous vars. pick an appropriate # of bins.
For the discrete vars. it is important for the hist to have at least 2 bars with more than 1 count.
In R, if x=qualitative, e.g. x=c("a", "b", "c"), do plot(as.factor(x)) to make a histogram.

Keep a copy of the data set because you will need it for other hw problems while this hw is being graded.

Here is motivation for putting effort into this hw problem:
Throughout the quarter we will be applying a wide range of methods to this data set. In past quarters there have usually been a few students who manage to take all of that analysis and turn it into something more than just hw. There have been technical reports, conference posters, and even journal publications. One student even published it in the very prestigious journal Science. So, if you're interested in something like that, let me know and I can help you out. So, it's best to put in some effort at the beginning to collect "good data." Every hw related to this data will be an opportunity to see if the data is good, and you can always change and/or update your data throughout the quarter.