Lecture 2 (CR.1)

Last time:

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. I

"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 a population from a single sample.</td>
</tr>
<tr>
<td>median</td>
<td></td>
</tr>
<tr>
<td>mode</td>
<td></td>
</tr>
<tr>
<td>range</td>
<td></td>
</tr>
<tr>
<td>histogram</td>
<td></td>
</tr>
<tr>
<td>scatterplot</td>
<td></td>
</tr>
</tbody>
</table>

E.g. From a sample, we can compute sample mean, sample range, sample....

What do these say about the population mean?
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** \( x \in \mathbb{R} \)

   e.g. \( x = \) time it takes to complete a computer code.

   b) **Discrete** \( x \in \text{Integers} \)

   \( x = \# \text{ of defective elements in a computer} \), \( x \in \{0,1,2,\ldots\} \)

   \( x = \# \text{ of Macs in a class of 100 students} \), \( x \in \{0,1,\ldots,100\} \)

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:

time to run some code.

<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>

Note: If the values did not change, we would not call it data! i.e. change is very important in data.

Infact a lot of statistics is about quantifying and understanding that change. The word “Variance” will come up a lot in this class.

This categorization may seem easy, but it’s NOT. Look:

Here are some ambiguities:

- Is $X_2 * 10,000$ discrete?!

Answer: It depends.

- Suppose you observe $X_2$ 100 times, but got

$$1.13, \ldots, 2.21, \ldots, 1.67, \ldots, 0.51, \ldots$$

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 con.

What’s the cutoff/boundary between discrete and con?

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 between variable types matters is in making histograms.

For discrete & categorical r.v.s, histograms are easy to make: just count the # of cases for each level of the variable. E.g. \( x = \) "favorite fruit type" \( \rightarrow \) data on, say, 7 people.

\[ x = \text{Orange, Apple, Banana, Orange, Kiwi, Orange, Apple} \]

The histogram is constructed by counting the frequency or count of observations falling into each category. If \( x \) is qualitative, then their order is arbitrary. Then the shape of the histogram is also arbitrary.

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:

\[ \text{hist}(x, \text{breaks} = \ldots) \]

\textit{breaks} controls the number of bins, approximately.

See lab 1.
The shape is important! But,

\[
\begin{align*}
\text{small bin size} & \Rightarrow \text{bunch of short bars scattered across x-axis.} & \text{Either way, useless!} \\
\text{large bin size} & \Rightarrow \text{few large bars.} & \text{...}
\end{align*}
\]

In lab you learn how to “turn the knob” on bin size to reveal hidden patterns (e.g., the existence of 2 different groups).

[HW - lect 2-1] Come-up with 2 examples for each of the above 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. Specifications:

1) number of cases: 30 or more

2) 2 categorical/discrete variables. One of them can 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) 2 continuous variables.

4) The 4 variables must relate to a common problem; not 4 unrelated variables.

a) print the data in the following format, and turn it in:

```
Variable 1    Variable 2    Variable 3    Var. 4
```

b) plot histograms for each of the 4 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", "a", "c") do plot(as.factor(x)) to make histograms.

Keep a copy of the data set because you will need it for other hw problems.
This document was created with Win2PDF available at http://www.win2pdf.com. The unregistered version of Win2PDF is for evaluation or non-commercial use only. This page will not be added after purchasing Win2PDF.