Teaching Intro Stats

Outline

What is statistics?

UNIVERSITY OF TORONTO MISSISSAUGA

Definition

Statistics is the science of data. This involves collecting, organizing, summarizing, analyzing, presenting and interpreting numerical and categorical information.

Two areas of statistics

There two primary areas of statistics:

1. descriptive statistics
2. inferential statistics

Loosely speaking, descriptive statistics involves describing data, whether it be numerically or graphically. Inferential statistics is the science of inferring about a population based on a sample from this population.

Topics for an intro stats course

There are several core topics in most intro statistics courses:

1. examining variables and data
2. understanding relationships between variables
3. probability
4. inference
5. regression

Some other topics

In addition to those five topics, time permitting, two other topics sometimes covered are:

1. sampling methods
2. design of experiments

These two topics are almost completely conceptual, with a lot of definitions.

What the students easily grasp...

Intro stats classes typically start with exploratory data analysis.

This would cover types of variables and examining data on such variables.

The students are able to make simple plots such as bar plots, histograms and pie charts, and furthermore, interpret them well.

What the students easily grasp...

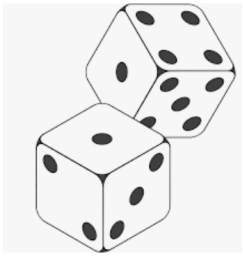
Students are also typically comfortable doing basic calculations and numerically describe data through such measures as mean and median.

They can understand concepts like percentiles and measures of spread.

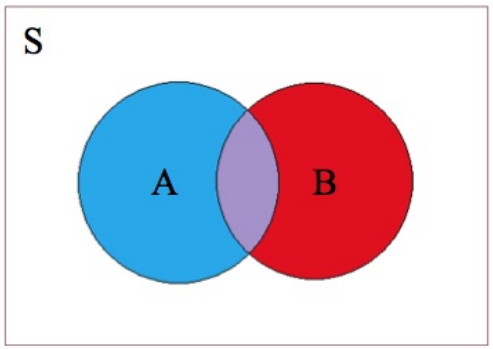
Even when two variables are examined together, they usually can grasp the concept of a relationship, as well as correlation.

What the students easily grasp...

Probability is another topic most students do well in, especially since it is easy to come up with simple examples they can relate to.



Using such things as Venn diagrams and tree diagrams, they can do basic calculations and with the help of further visual tools and simulation, they seem to do enjoy the topic.



What the students struggle with...

Across several of the topics in an intro stats class, some basic math skills are required, and this is where I often start to see some students struggle.

For instance, in random variables and distributions, as well as in regresion and inference, some basic calculations cause students a lot of anxiety.

For example, solving an equation in one unknown is one of the first hurdles I see students face.

What the students struggle with...

Consider the topic of probability mass functions (PMFs); they may be asked to find the value k which makes the following function serve as a PMF for a discrete random variable X:

$$p_X(x) = \begin{cases} k(x^2+4) & \text{if } x = 0,1,2,3, \\ 0 & \text{otherwise.} \end{cases}$$

or probability density curves (PDFs) where they may be asked to find the value c that makes the following a valid PDF for some variable X:

$$f_X(x) = \begin{cases} cx(1-x) & \text{if } 0 \leq x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

What the students struggle with...

They also struggle with symbols, such as Σ and $\Sigma_{i=1}^{n}$.

They often don't understand when we write $\sum_{i=1}^{n} x_i$.

What the students struggle with...

In the past couple years, more software is being introduced into courses, and not the typical drop-down menu software like Excel.

Typically, courses are now using R and unless they have a computer science background, many are seeing coding for the first time and are terrified of it!

Some new teaching approaches

Over the past few years, myself and my colleagues have been trying to modernize (more traditional/theoretical) stats courses and one of the ways we've been doing this is in fact introducing statistical software.

This means, for instance as I mentioned, the software R is now being (gently) used in intro stats classes that typically used software like Excel, or Minitab (which are very user friendly and in general, are non-coding and more drop-down menu).

Introducing simulation

One area I'm particularly keen on introducing software is the area of probability, and namely through simulations.

Typically, simulation is somethings would only see in a higher level statistical computing class, but I am currently redesigning an intro stats course at UTM where will I incorporate simulations, which to my knowledge, has not been done in this course thus far.

For example, I plan to demonstrate simulating simple random experiments, such as tossing a coin or rolling a die....

Intro stats and data science

We all know that data science borrows concepts and methods from mathematics, statistics and computer science.

The driving themes behind data science are collecting, anlayzing, visualizing and interpreting data.

Looking at intro stats courses, how can we teach intro stats and data science to first year students?

Introduction to statistics and computing

In a data science program, from a statistics perspective, the first course a student would take is Introduction to Statistical Computing.

Such a course will introduce all the fundamental concepts you'd see in an intro stats class plus computing.

The computing would be done in a statistical software package, like R.

Officially I don't think UTM has such a course, but I hope to teach it!

Thank-you to Andie for inviting me and thank you all for your attention!