Lecture 2 (Ch. 1)

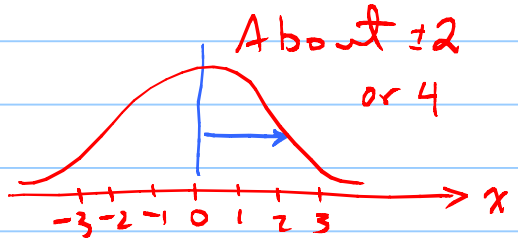
Last time

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. 2

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



Two Types of Statistics:

Descriptive

mean
median
mode
range
histogram
scatterplot

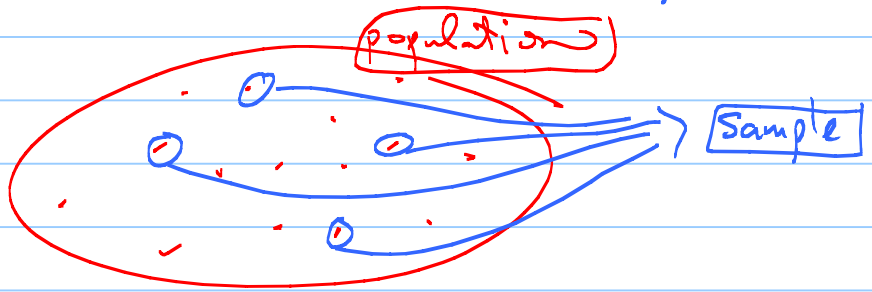
Inferential

To infer something about a **population** = Technical term
= A set (finite or not) to which we have incomplete access.

from a **Sample**

= Technical term

= A subset to which we have complete access.



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 diff. 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, on a random day.

b) Discrete $x \in$ Integers (see qualification, next page).

x = # of defective elements in a random computer. $x \in \{0, 1, 2, \ldots\}$

x = # of Macs in a random class of 100 random students $x \in \{0, 1, \ldots, 100\}$

2) Qualitative (or Categorical)

x = computer type in a random class. $x \in \{$Mac, Dell, HP$\}$

x = state of a random coin. $x \in \{$Heads, Tails$\}$

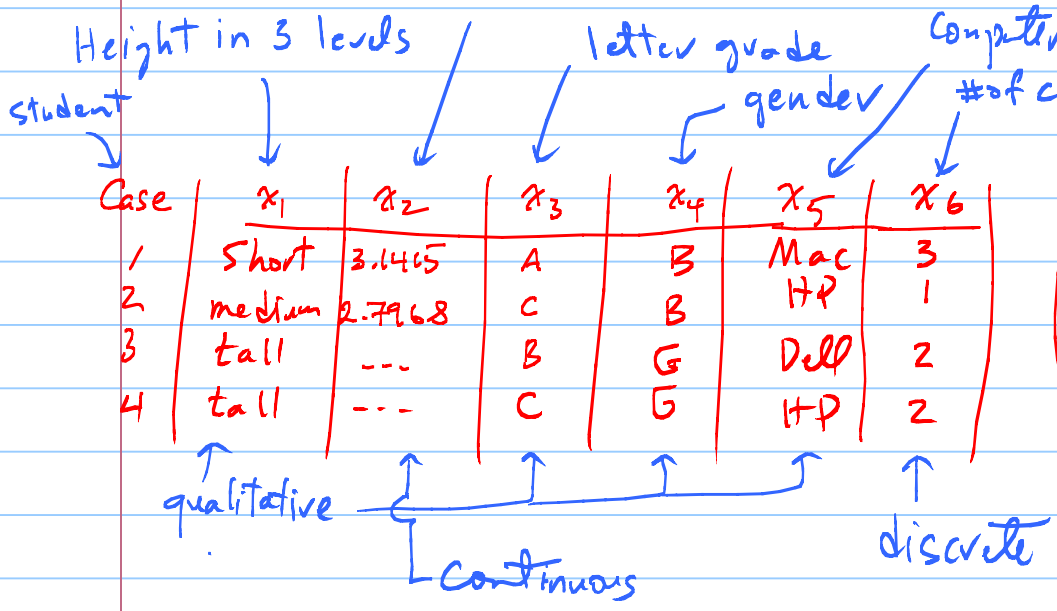
x = Letter grades in a random class of 120 random students $x \in \{A, B, C, D, F\}$

called "level"

Random Variable. This is a very important (but dense) concept in statistics and data analysis. All we need to know about it is that it is a variable (e.g. length, time, fruit type) that changes values every time we observe/measure it. So, when you measure your weight several times, and get several different values, that makes "weight" a random variable.
There are different reasons for why things are random; in the weight example, it's because your weight scale is not perfect. In other situations it's because the observations we make are on a random sample taken from the population. For example, if we consider a stat 390 class as a random sample of size 120 taken from the population of all students who have taken (or will take) stat 390 students, then the "class mean grade" is a random variable. By contrast, the mean grade of all students who have taken (or will take) stat 390 is NOT a random variable. That mean is not subject to variability at all; it's a unique number that exists, but we don't have access to it. For that reason, it's called the population mean; recall the defn of population. The other mean, i.e., of a class is called the sample mean. It's the sample mean that's a random variable. The population mean is not - it's called a population parameter. We'll talk more about these later. But, for now, the take-away should be that things dealing with (random) samples are random variables, while things pertaining to the population are not.

So, the word "random" is implicit in many places in the above examples.

Data (ie. Sample) on These r.v.'s may look like This:



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

The differences between These types may seem straightforwad, BUT IT IS NOT!

Here are some ambiguities:

- Is $x_2 * 10{,}000$ discrete?!

Answer: It depends.

- Suppose you observe x_2 100 times, but get

1.13, ..., 2.21 ..., 1.67 ..., 0.51 ...

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, and messy.

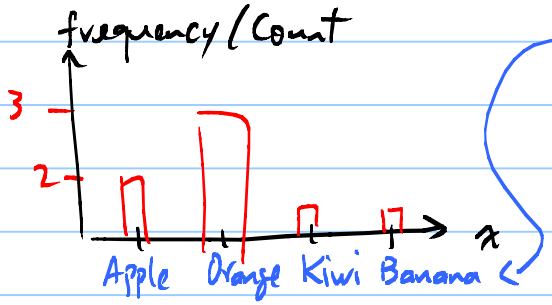
Really, really, powerful data analysis tool!

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.

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

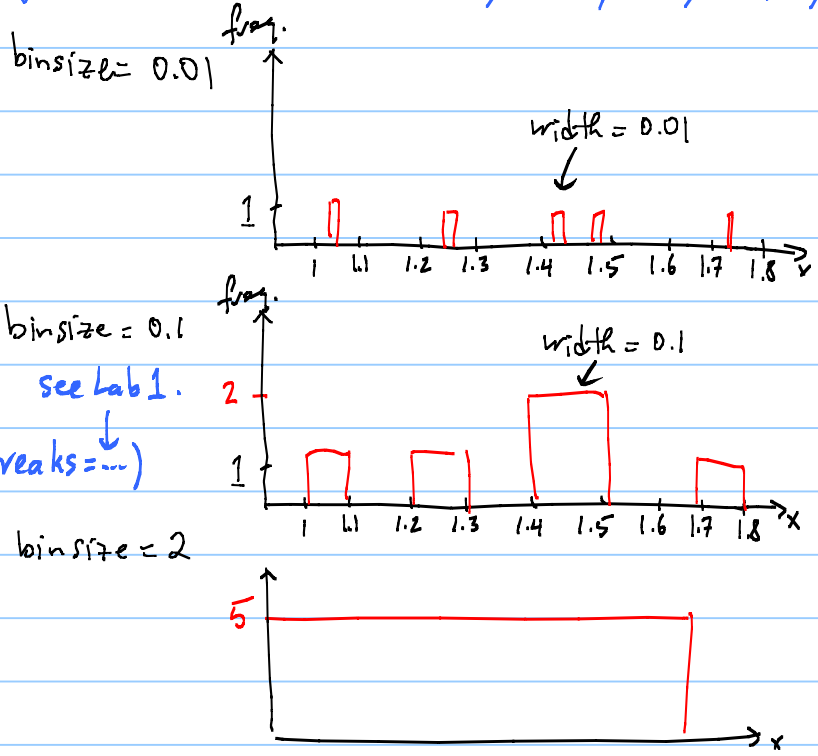


→ If x = qualitative, then their order is arbitrary. Then the shape of the hist is also arbitrary.

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: see Lab 1.
hist(x, breaks = ...)

⇒ bin too small (useless)

In Lab you learn how to "turn the knob" to reveal hidden patterns in data (e.g. the existence of 2 groups.)
shape is important.

⇒ bin too large (useless)

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

hw_lect2_1

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, ie. the type depends on the specifics of data. Here, however, ignore that complexity, and base your answer on theoretical considerations (ie. based on what you know about that variable).

hw_lect2_2

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 (the more, the better),
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 first 30 cases, in the following tabular 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