Lecture 2 (Ch. 1)

Last time we got to the concept of a random variable (r.v.).
We also talked about different kinds of r.v.s.
Let me refine those defns.

At the highest level we have quantitative vs. qualitative.

These distinctions matter, because each type has a different methodology developed for it.
In 390 the methods we learn care only about whether the r.v. is continuous or discrete/categorical.

Data (ie. sample) on these r.v.'s may look like this:

time to run some code.
Height in 3 levels
letter grade
gender
computer Brand.
number of computers

Case	x_1	x_2	x_3	x_4	x_5	x_6
1	short	3.1415	A	B	Mac	13
2	medium	2.7968	C	B	HP	10
3	tall	---	B	G	Dell	10
4	tall	---	C	G	HP	11

qualitative — continuous — quantitative.

There are other "finer" types of variables, too: ordinal/nominal/---
But we don't deal with them in this course.

Here is one ambiguity: is $x_2 * 10{,}000$ discrete?!
Answer: It depends. Read on!

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 got 50 distinct/different values, then treat it as cont.

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

Answer: It depends on lots of things, e.g. the total sample size.
And/or what you want to do with the data.

You will gain some experience in this class.

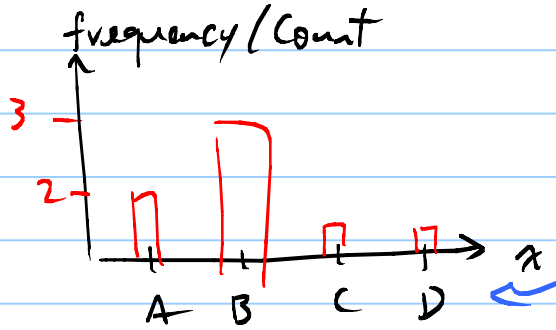
One place where the difference between cont. and categ is important is in assessing the "distribution" (in the lay sense of the word) of data. In statistis, when we want to talk about that distr. (ie. "the distr. of data"), we call it a histogram.

A distribution is something else! Later.

For discrete / Categorical, They are easy to make,
Just count The # of cases for each level of The variable.

Eg. $x = A, B, B, C, A, B, D$

The hist is



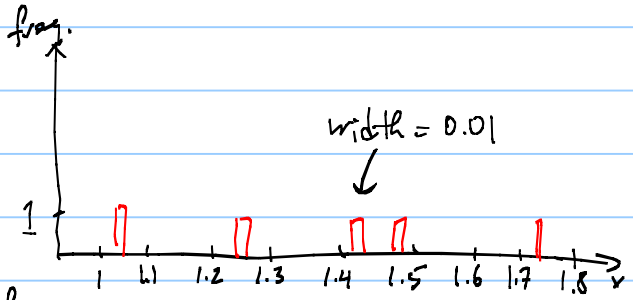
If x = qualitative, Then Their order is arbitrary. Then The shape of The hist is also arbitrary.

Histogram for continuous x :

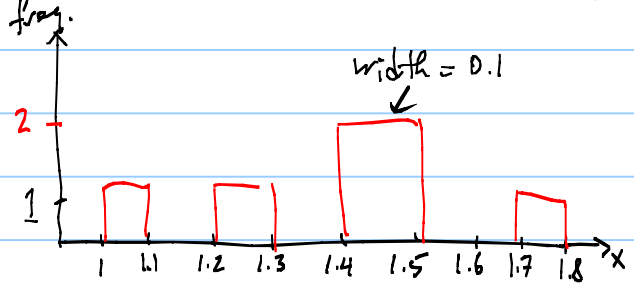
Divide-up The x-axis into some number of intervals/bins, and count how many cases fall in each bin / interval.

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

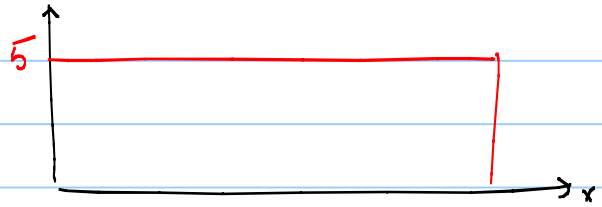
binsize = 0.01



binsize = 0.1



binsize = 2



In R:

hist(x, breaks = ...)

controls The number of bins, approximately

See lab 1.

The shape is important! But.

small binsize ⟹ bunch of short bars scattered across the x-axis.

No good either way!

large binsize " ⟹ few large blocks.

In Lab you learn how to "turn the knob" that controls the bin size (or their number) i.e. "breaks" in R. revealing useful info, e.g., 2 different groups.

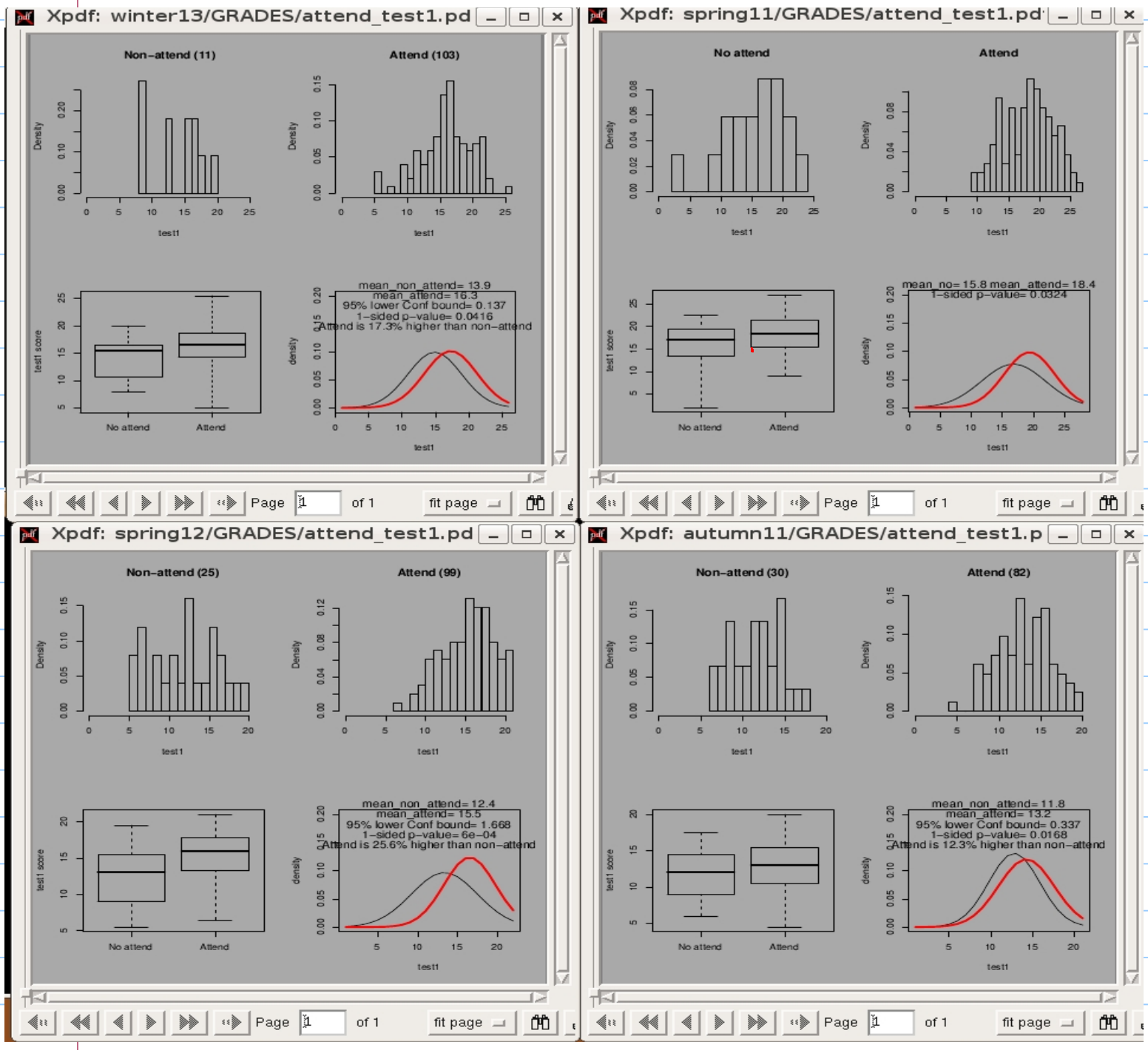
Important.

There is a great deal of useful info in a histogram:
e.g. center (location) of data = typical value
spread of data, = typical spread of values
shape of data, ... All tell a good story.

⇒ In the future, the first thing you should do when you see a bunch (a column) of observations (either numbers or not) histogram them. You will learn something! ⇐

The interpretation of histograms is an "art" that you learn through practice. Here is one example:

Here is an example use of a histogram that should interest all of you. Just concentrate on the hists; you will learn about the rest, later.



All of this suggests that attending 390 lectures is associated with higher test grades. This is from only 4 quarters, but the same pattern exists for every quarter! Of course things may not be causal.

A Huge and Tricky concept

We have been talking about data, and histograms of data.

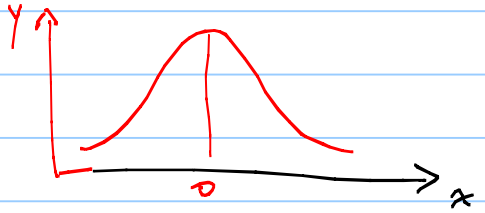
A histogram pertains to data.

But there is something else that looks like a histogram, but it's not: A DISTRIBUTION

A dist. is a purely mathematical thing that has nothing to do with data. So, for now, forget data (and hists).

In statistics, distributions are used to represent the population, while histograms are used to describe the sample (data). Later, we are going to learn how to tell something about the pop. (ie. distr.) from a sample (ie. histogram). But, again, dists and hists are completely different things.

Example: $y \sim f(x) \sim e^{-\frac{1}{2}x^2}$



Technically, this f(x) is not a distribution! You will see why, tomorrow. But it's good enough to make the important point that a distr. is a mathematical thing (ie. a function), not a histogram, even though they look alike.