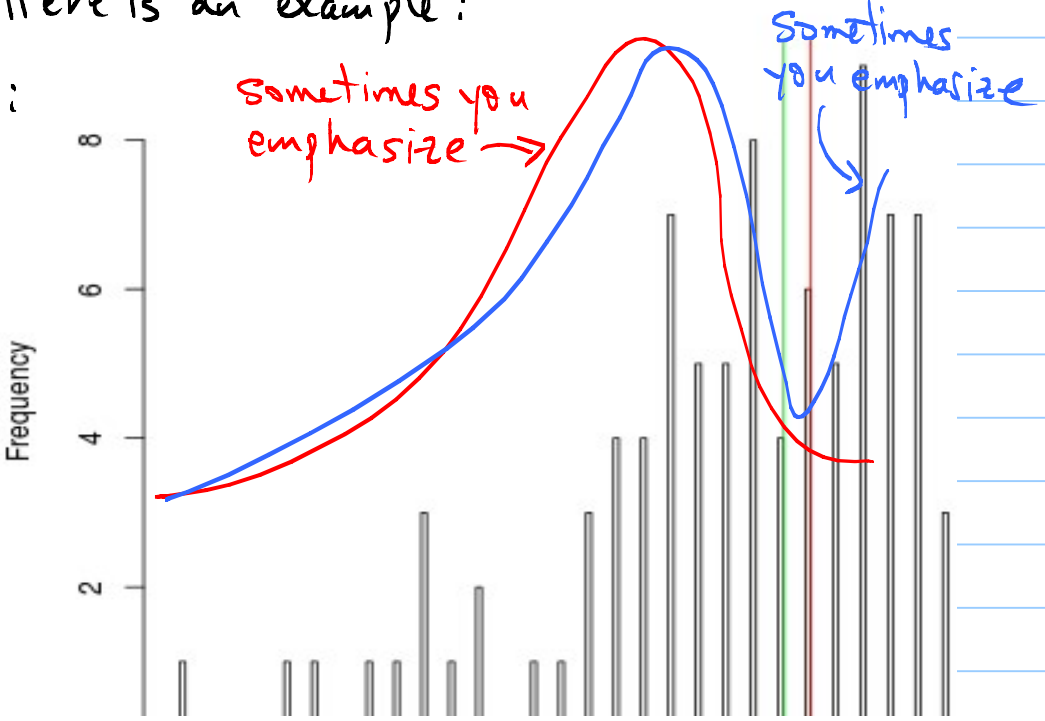
Lecture 3 (Ch. 1)

We've talking about histograms; very useful for data analysis.

The shape of a histogram is very important, and conveys lots of info.

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

Grades from Spring 17:



Ignore the "small" features. See the "big picture"

What's small? What's large?

Ans. It depends!

There exists a great deal of information in the shape of a hist.

Additionaly, 2 summary quantities are

- center (location) of data = typical value in data
- spread (width) of data = typical deviation/spread in data

The word "typical" is important when we interpret summary measures

⇒ In the future, the first thing you should do when you
⇒ see a bunch of observations (either numbers or not)
⇒ histogram them.
⇒ Then interpret (at least) 3 things: shape, center, spread
⇒ You will learn something!

prob
prop

Histograms can show frequency, or relative freq. on the y-axis:

Rel. freq. = freq. / total sample size. (examples on next page).

For rel. freq. hists:

If x = discrete/categ : (height of bar at $x=a$) = $pr(x=a)$

If x = contin : (height of bar at some bin) $\propto$ (proportional to) $pr(x \in \text{bin})$

(Not equal (see below).)

Here, "pr" stands for either "proportion of times" or "probability"

Note that we have introduced a very important connection between hists and probs. But that connection depends on the type of variable:

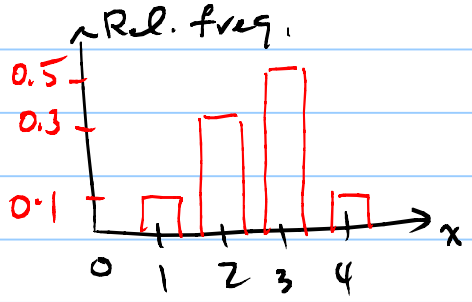
Examples:

Suppose x = discrete with this hist:

$pr(x=2) = 0.3$

$pr(2 \le x \le 3) = 0.8$

$pr(x \le 3) = 0.9$



Suppose x = cont. with the same hist:

Then pr(--) will depend on bin size; and to account for that some hists show "density," i.e. (rel. freq. / bin size). Either way, suffice it to say that for x = cont., we get forced to look at areas (not heights) as probs. Consequently, for x = cont.

$pr(x=b) = 0$. $\Rightarrow$ It also follows $pr(x \le b) = pr(x < b)$

Compare

This difference between cont. & disc/categ. variables, in terms of how they give probs, will keep showing up in this class. So, watch out!

Example of rel. freq. hist and their interpretation:

Caren Marzban Other SP16

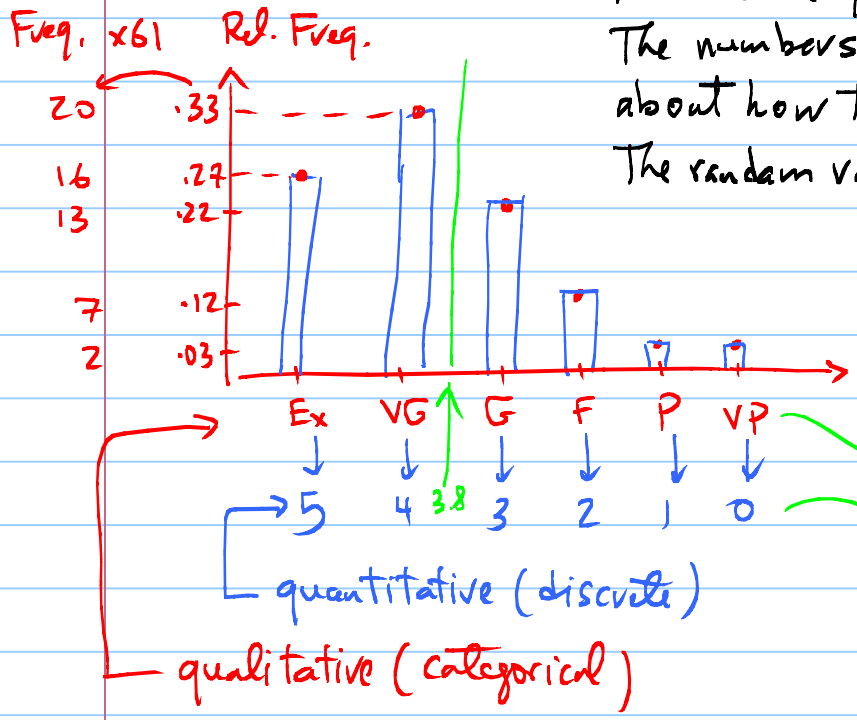
sample "61" surveyed population "124" enrolled

Form G: Lecture -- Assignments

Question	Excellent	Very Good	Good	Fair	Poor	Very Poor	Median
The course as a whole:	27%	33%	22%	12%	3%	3%	3.80
Textbook overall:	33%	30%	27%	10%	0%	0%	3.94
Instructor overall:	50%	28%	10%	7%	2%	3%	4.50
Instructor's contribution:	42%	27%	15%	8%	3%	3%	4.22
Instuctor's interest:	53%	26%	7%	5%	2%	7%	4.56
Amount learned:	39%	27%	20%	8%	3%	2%	4.09
Relevance and usefulness of homework:	37%	17%	27%	12%	3%	3%	3.75

There are many more of these at the bottom of course website.

For median calculation: 5 = Excellent 4 = Very Good 3 = Good 2 = Fair 1 = Poor 0 = Very Poor



Make sure you identify the r.v. on the x-axis. The numbers on each row say something about how the students rated something. So, the random variable (r.v.) is Rating.

median ? #!

median = 3.80

median (and mean) make sense only for quantitative data.

quantitative (discrete)

qualitative (categorical)

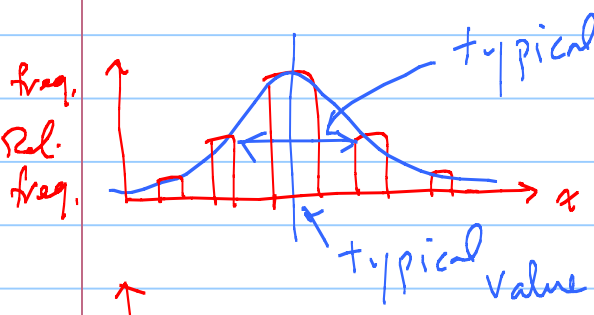
Interpret: center ~ 3 or 4
spread ~ 1, 1.5
shape ~ skewed (to ...)

Also, can find prob(rating = ...)

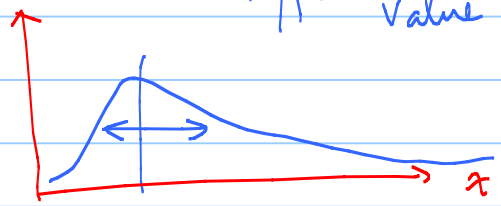
Types of hists

Here are some of the special shapes that you may come across:

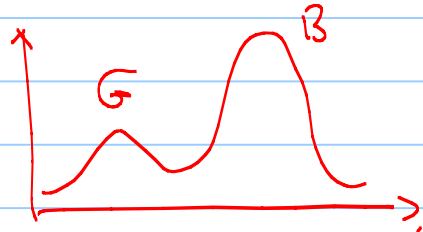
and their interpretations



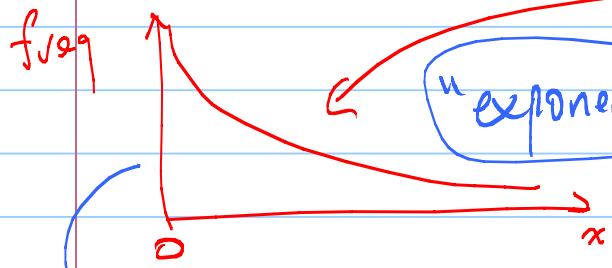
bell-shaped



skewed



bimodal



"exponential-looking"

A hist is a plot of freq. (or rel. freq, ...) of different values of ONE variable.

NOT some variable as a function of time!

NOT some variable (e.g. demand) as a function of some other variable (e.g. supply).

FYI

It turns out there are 2 special hists both of which look exponential:

$freq = e^{-\lambda x}$ — exponential

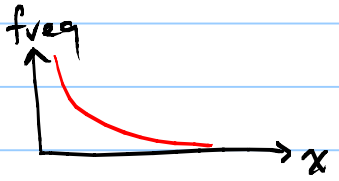
$freq = x^{-\lambda}$ — power-law

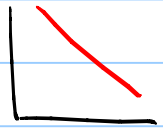
next page shows how to find out which one you may have.

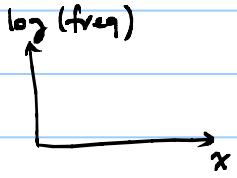
In addition to hist(x) it's also useful to look at hist(log(x)). Sometimes, if hist(x) is skewed, hist(log x) will be bell-shaped. Usually, looking at hist of log(x), or $\sqrt{x}$, or some other transformation of your data, will make the hist more bell-shaped, and that's a good thing, because of easier interpretation and easier Math (later).

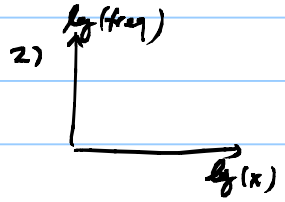
FYI

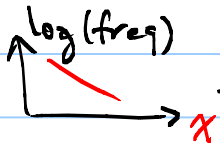
When you get an exponential looking hist, the way to determine whether you have an exponential or a power-law histogram is to transform:

If you get 

, Then look for  in

1) 

or
2) 

1) If you get 

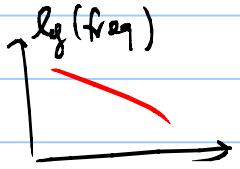
, Then the freq. is proportional to

$\log(\text{freq}) = \alpha - \beta x \Rightarrow \text{freq} = e^{\alpha} e^{-\beta x} = (\text{constant})\, e^{-\beta x}$

ie. The frequency hist. is really exponential.

As a result, the freq. hist is called exponential. (More later)

In short, $\text{freq}(x) \sim e^{-\lambda x} \iff \log(\text{freq}(x)) \sim -\lambda x$

2) If you get 

, Then

$\log(\text{freq}) = \alpha - \beta \log(x) \Rightarrow \text{freq} = e^{\alpha} e^{-\beta \log x} = e^{\alpha} x^{-\beta}$

These hists are said to follow a "power-law". E.g.

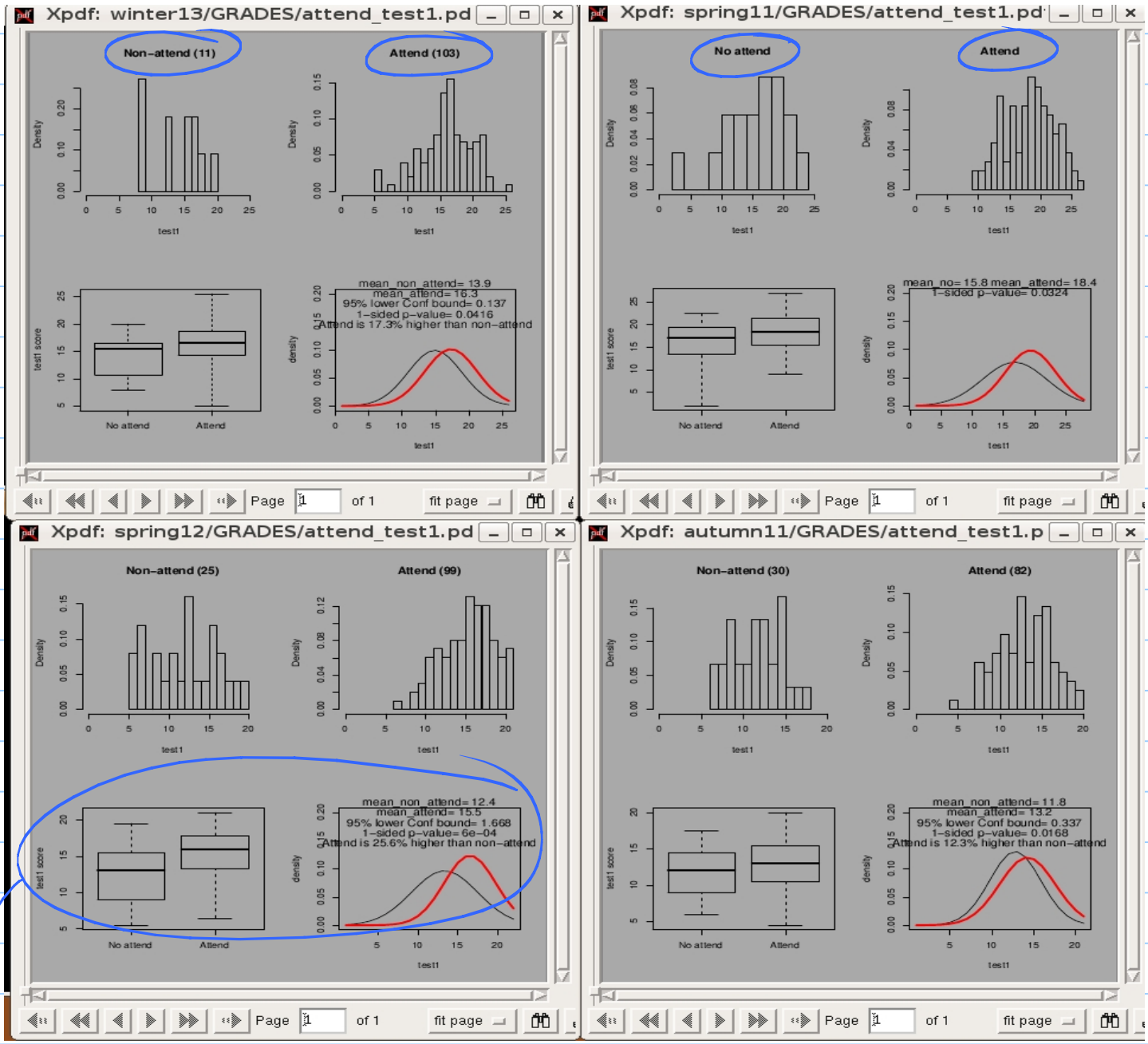
x = magnitude of earthquakes
= population of cities, on the planet
= length of words, in a book
= casualties of wars, for different wars

R: hist(log(x)) ≠ plot(log(hist(x)$mids), hist(x)$counts)

Because hist(log(x)) gives freq(log(x)) not log(freq(x)).

Here is another example of the use of histograms, showing that attending lectures leads to higher grades.

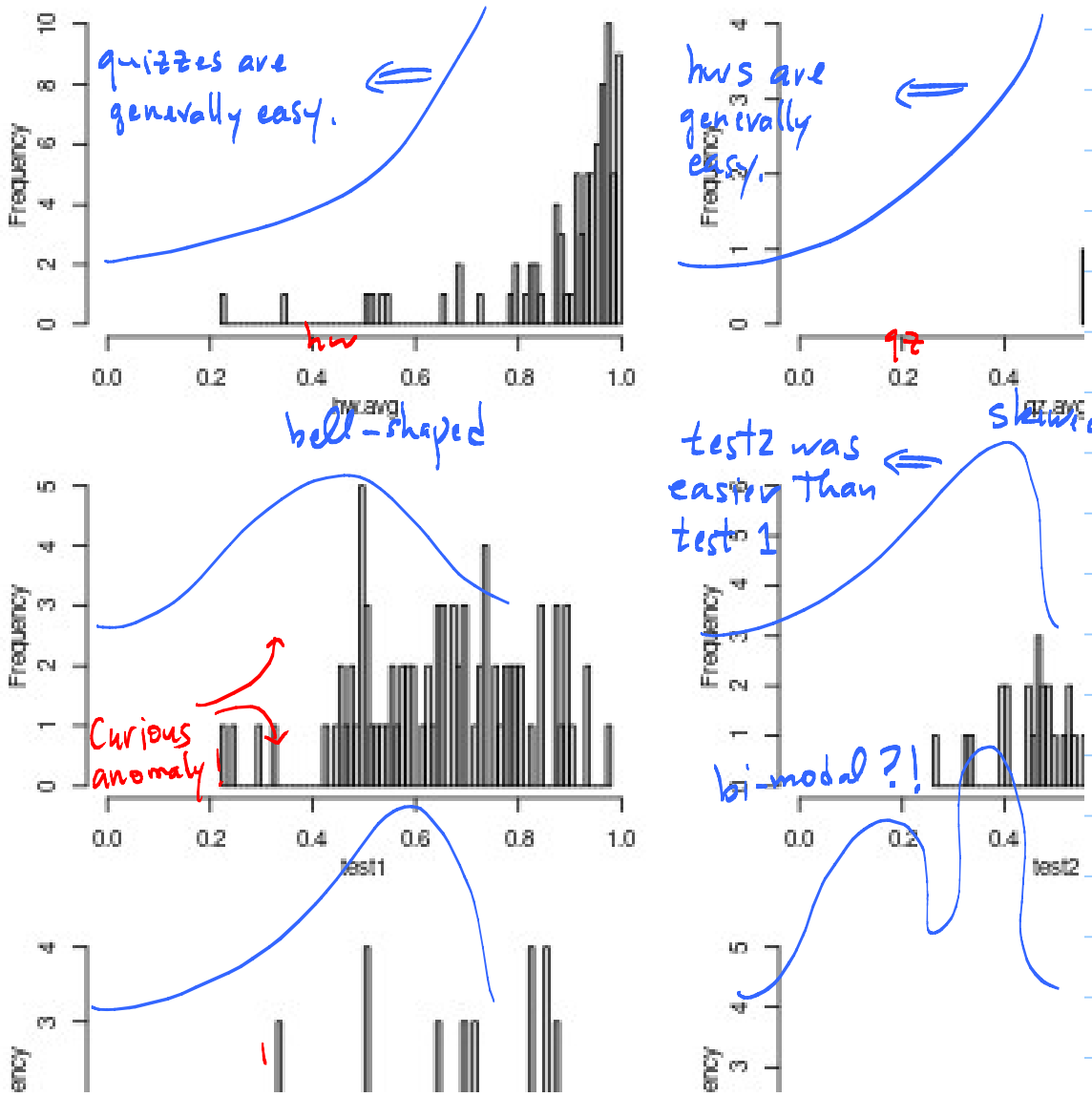
Histograms of test grades for non-attending and attending students.



You will learn about the rest of this plot throughout this class.

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.

Another example (from autumn 20) where many shapes show-up. More importantly, learn to interpret them "In English". E.g.



hw_lect3_1

For each of the following shapes, come-up with at least one example of a random variable x (continuous or discrete) whose histogram you expect to be approximately

a) Bell-shaped (symmetric)

b) Skewed (one way or the other)

c) Exponential-looking

d) Bi-modal

Make sure you describe/define the random variable clearly (like we did in the lecture), and explain in words why you expect the particular shape. If you have data to support your expectation, then show the histogram.

hw_lect3_2

In this lecture there are many examples of random variables that, when considered as quantitative, have an exponential-looking histogram. Identify one of the random variables, and plot its relative frequency histogram. Hint: The relative frequencies are, in fact, in this lecture, too.