Lecture 2 (Ch. 1)

So far, continuous vs. discrete (or Categ.) $\begin{cases} \text{quantitative} \\ \text{qualitative} \end{cases}$ Data

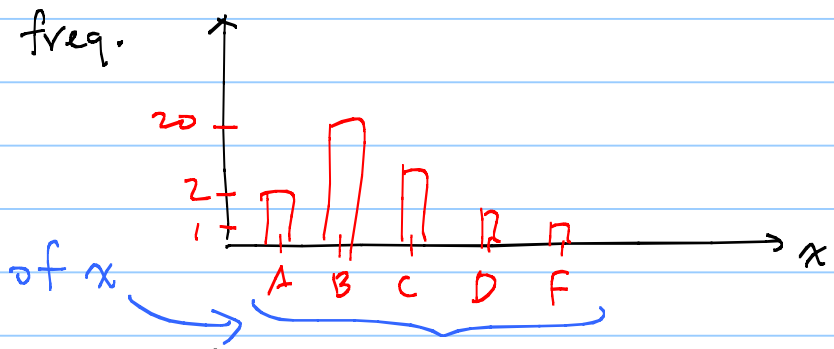
One place where the difference between cont. and categ. is important is in assessing the "distribution" (in the lay sense of the word) of data. When we talk about data, we do NOT use the word distribution, but rather histogram

Dotplots and stem-and-leaf are alternatives.
Read about them, but we will do only histograms.

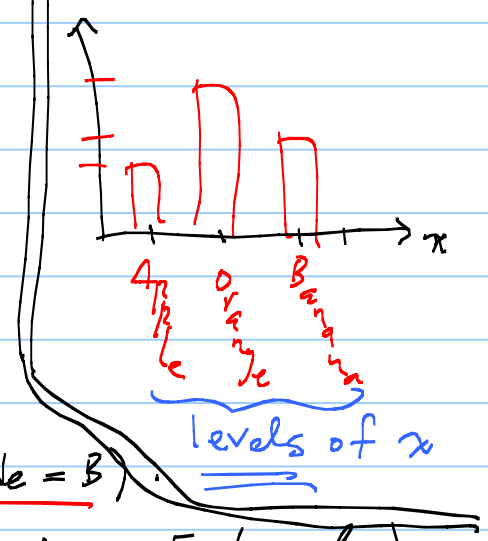
Histogram for discrete x: (a lot more in Lab)

Easy! Just count the occurrence of each level of the discrete variable, and plot that count (frequency) on y-axis.

E.g. x = grade (levels A, B, C, D, F)



x = fruit type



Conclusion from this histogram:

- The most frequent grade is B (ie. mode = B).
- There are only 5 distinct grades (ie. 5 levels)
- The minimum (maximum) grade is F (A). ← spread
- The histogram is not symmetric,
- ...

A realistic example:
My student evaluations from a past quarter:

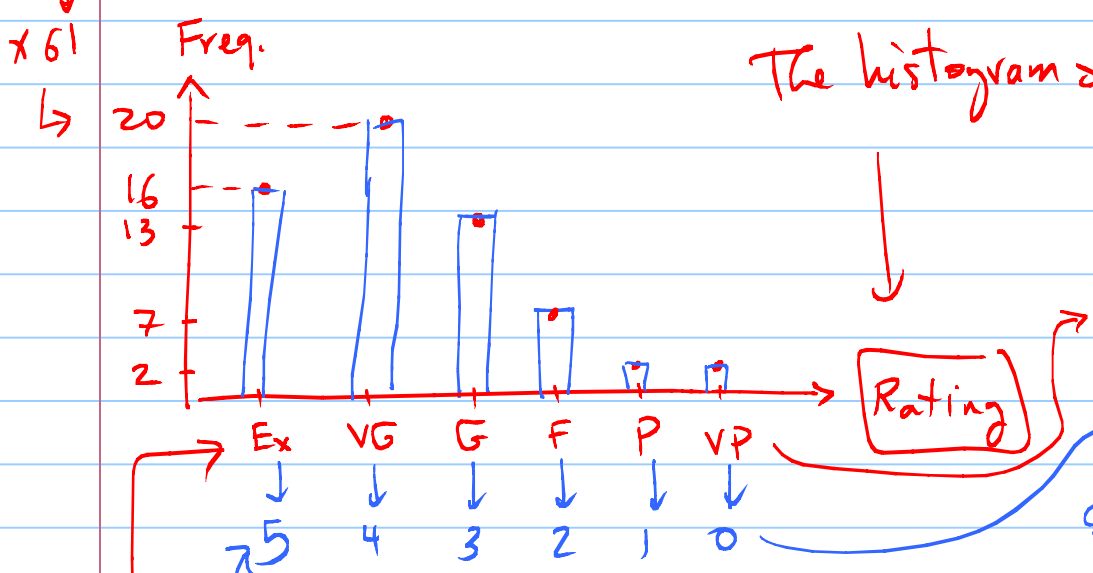
Caren Marzban Other SP16

Form G: Lecture -- Assignments "61" surveyed "124" enrolled

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

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

The histogram of rating (or quality).



median ? #!

median = 3.80

So, median (and mean) make sense only for quantitative data, not qualitative data. Even, then one still needs other info to compute median. Infact I don't know how to get 3.8!

Categorical (but quantitative.)

Categorical (but qualitative)

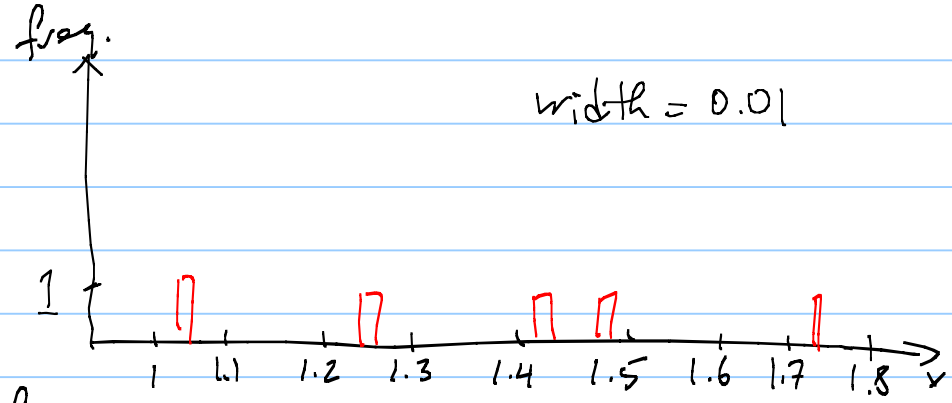
Histogram for continuous x :

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

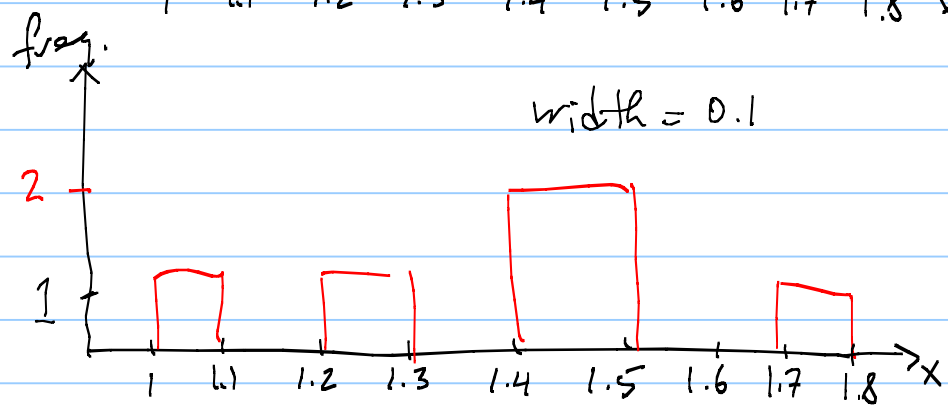
Eg. Data: $x = 1.05, 1.25, 1.41, 1.48, 1.75$

binsize = 0.01

too small

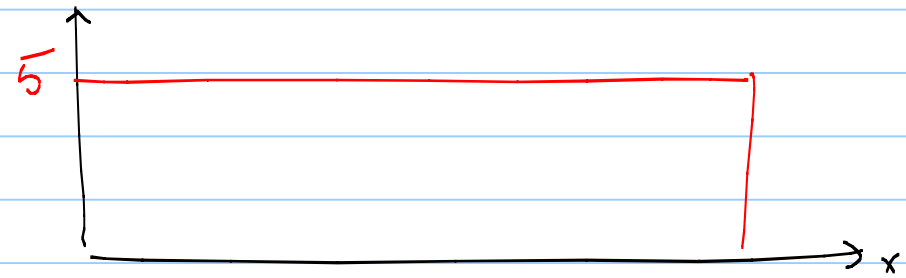


binsize = 0.1



binsize = 2

too large



in R : hist(x, breaks = ...) ← in Lab.

Controls the <u>number</u> of bins.

The point: The shape of histogram is useful
But its shape does depend on binsize

small binsize $\Rightarrow$ bunch of short bars scattered across the x-axis.

No good either way!

large binsize " $\Rightarrow$ 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.

Random Variables (Important Concept!)

Each variable labeling the x-axis of the histograms above is called a random variable. It's a variable that takes random values; the values are called Levels. Eg.

x = grade of 120 students, levels: A, B, C, ...
x = fruit type in a fruit basket, levels: orange, apple, ...
x = up-face of a coin, levels: H, T.

Two variations on histograms are Relative Frequency histograms and Density Scale histograms

→ Sometimes it's better to look at Relative freq histograms:

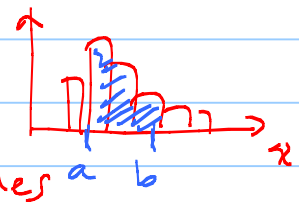
Rel. Freq. = Frequency / total sample size.

→ At times it's important to look at density scale histograms:

Rel. Freq. / bin size.

E.g. when binsize itself varies along x-axis.

Density histograms have 2 nice properties:



1) Area of each bar = rel. freq.

E.g. $\boxed{\text{area}}$ under hist. between two values

⇒ = $\boxed{\text{proportion}}$ of times x is between those two values.

2) Total area = 1 Good for probability (Later!)

Any/All of These histograms carry a great deal of information, and so, our lesson is, again

Moral: When you see a column of numbers, The 1st Thing you should do is histogram.

hw_lect2: In The above lecture note, There exists atleast one random variable That when considered as quantitative has a histogram That has no "hump," it does not look like  hump

Identify one of Them, and plot its rel. freq. hist. (By hand).