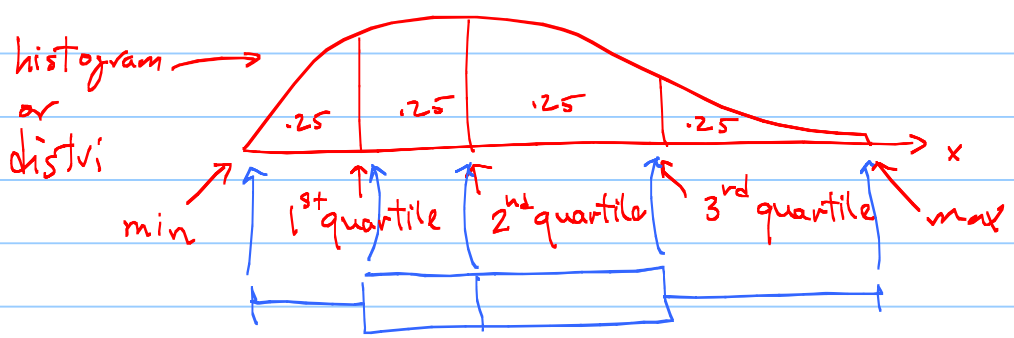
Lecture 7 (Ch. 1)

So far, for $N(\mu, \sigma)$, you can do

1) given x (or x's), find area
2) given area (eg. 90%), find x (or x's) ← percentile.

Percentile (or quantiles, quartiles, ...) apply to dist and hists.

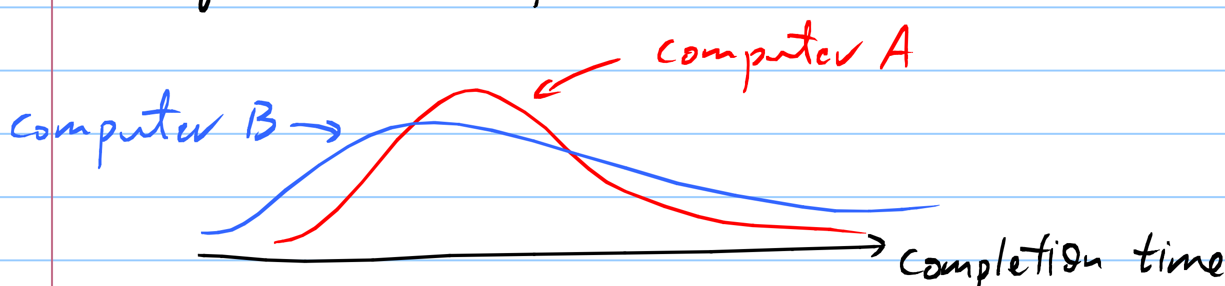
This material is from 2.3, but fits better here for lect. & lab.



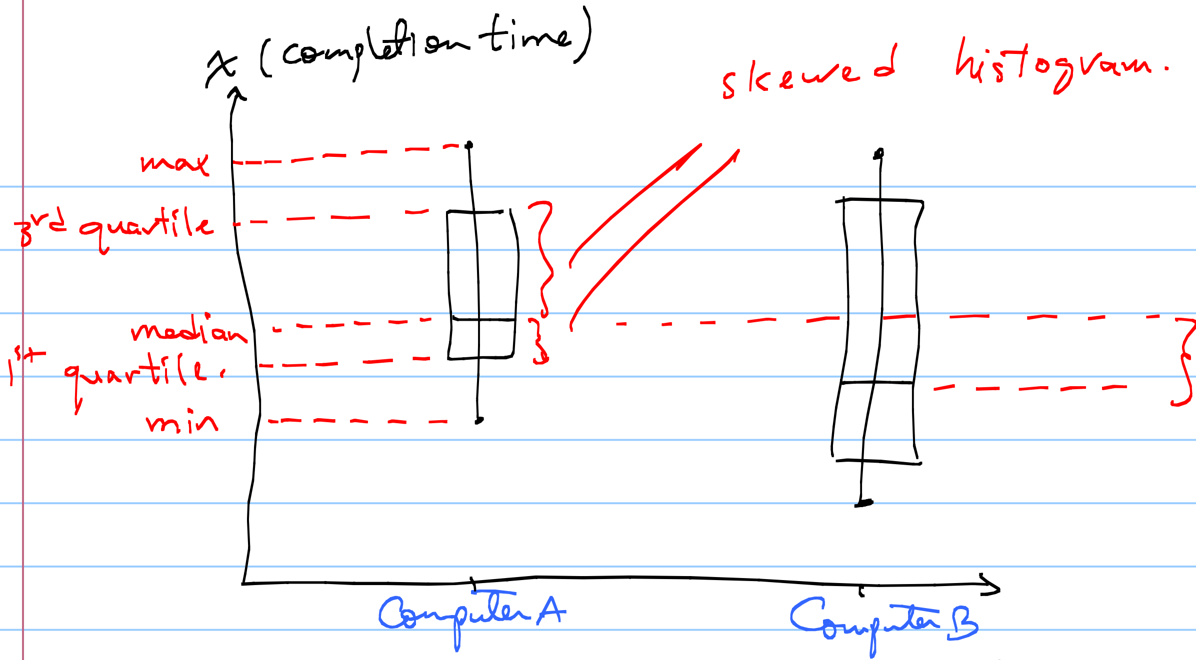
2nd quartile = 50th percentile = median = splits data in half.
1st (3rd) quartile = median of 1st (2nd) half.

Quartiles are the basis of of the so-called "5-number summary" of a hist (or dist), often plotted as a boxplot:

E.g. Suppose you want to find out which of two computers is faster. You take a given program, and run it on each computer 100 times, and record the time it takes to run the code to completion. You can then look at the histogram of "completion time" for the 2 computers:



The interpretation of such results is complex (see next page). Boxplots allow us to handle problems like this but involving many more (than 2 or 3) computers.



Observations: computer B is faster on the "average," because its median completion time is shorter. But computer B is also the more "moody" one (less consistent), because it has a wider spread in completion times. Important: Note spread !!

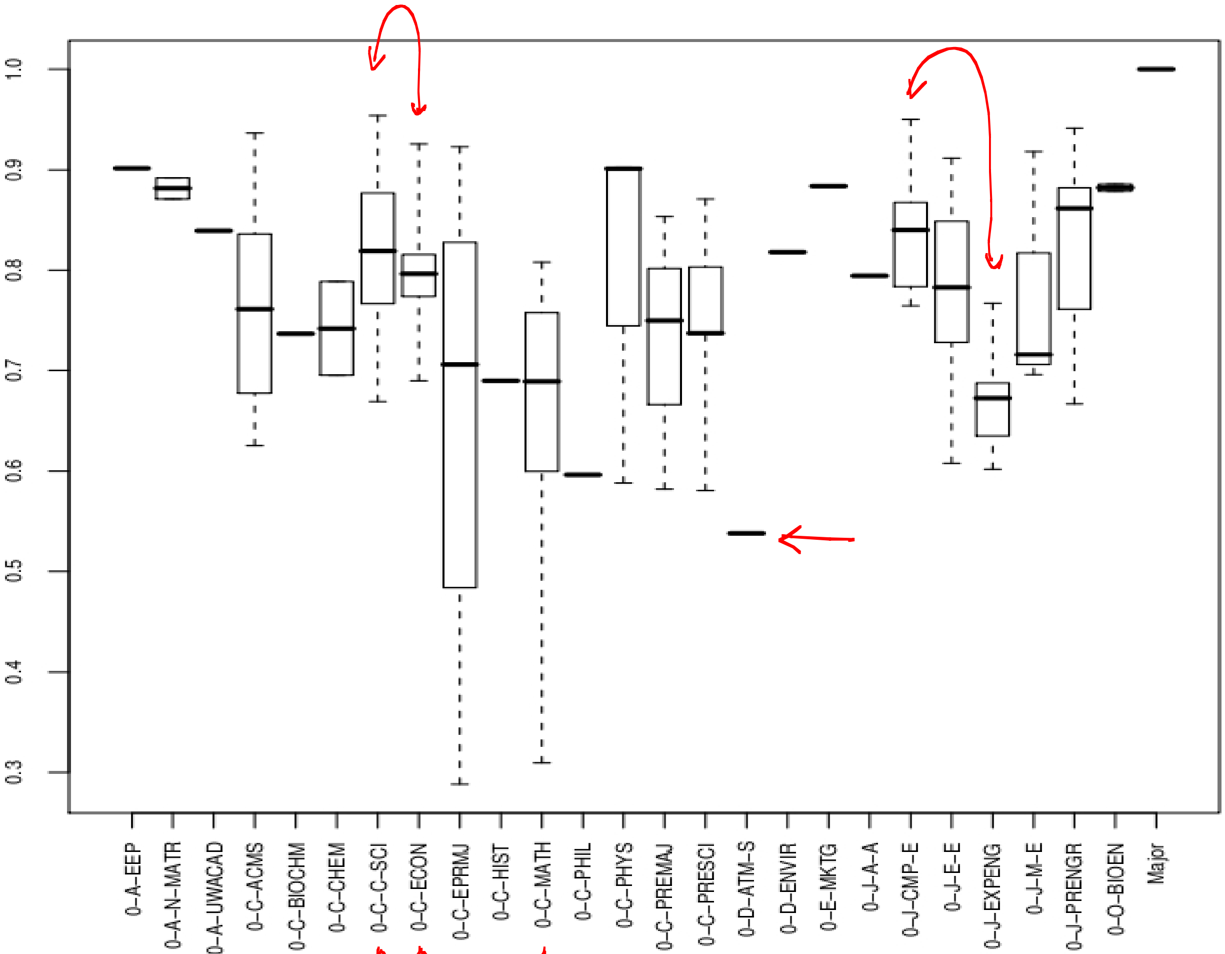
Having said all that, one cannot conclude that computer B is faster, because these boxplots are based on a <u>sample</u> of size 100. We do <u>not</u> know what the true distribution of x is. The true/population mean (or median) of x for each computer is somewhere in the boxplot, but we don't know where. Given the huge overlap between the boxplots, we can<u>not</u> conclude the B is faster. <u>We cannot conclude anything</u>! How much overlap is too much? Ans - in Ch. 7, 8.

For now, just learn that everytime you see a number, it's actually a sample (of size 1), and that it's actually a single realization of a random variable, and that the variable actually has a spread. And that's important!

Here is an example involving many groups (e.g. computers):

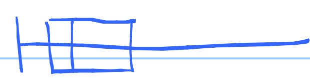
probably not different

probably different



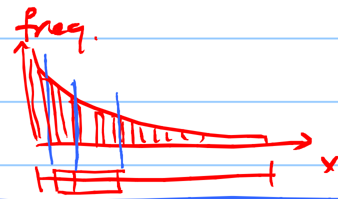
lots of class discussion!

Q1: Consider this boxplot of a histogram



The histogram itself will "look"

A) Uniform B) Normal C) Exponential D) cannot tell



In the above question, one can also conclude that the population (ie. distribution) from which the sample was drawn is probably exponential.

Recall that we use dists. to represent populations, and hists to represent the sample/data from that pop.

We have been using dists. as mathematical objects. And they are! But it may help to derive one

$\boxed{\text{Binomial:}}$

Consider N objects (population), where

Each object is 1 (Head, Girl, ...) or 0 (Tail, Boy, ...)

Suppose the proportion of 1's in the pop. is known $= \pi$.

Now, select n (e.g. 3) of the objects (with replacement) = Sample

and note the value of each object.

Repeat many many times (e.g. 10^8)

Q What proportion (of the 10^8) will be 1,1,1? 1,1,0? Etc.

Note: I'm not asking for the prop. of 1's in each sample.

I'm asking for the prop., out of the 10^8 trials, that are 1,1,1. Etc

A

		independence	X = # of 1's
prop. of	1,1,1	$= \pi \cdot \pi \cdot \pi$	3
	1,1,0	$= \pi \cdot \pi \cdot (1-\pi)$	2
	1,0,1	$= \pi \, (1-\pi) \, \pi$	2
	0,1,1	$= (1-\pi) \, \pi \, \pi$	2

$\left.\begin{matrix} 2 \\ 2 \\ 2 \end{matrix}\right\} 3$

Etc.

0,0,0	$= (1-\pi)(1-\pi)(1-\pi)$	0

$$\text{prop}(X=3) = 1 \, \pi^3 \qquad \frac{3!}{3!\,(3-3)!}$$

$$\text{prop}(X=2) = 3 \, \pi^2 (1-\pi) \qquad \frac{3!}{2!\,(3-2)!}$$

$$\text{prop}(X=1) = 3 \, (1-\pi)^2 \pi \qquad \frac{3!}{1!\,(3-1)!}$$

$$\text{prop}(X=0) = 1 \, (1-\pi)^3 \qquad \frac{3!}{0!\,(3-0)!}$$

$$\therefore \text{prop}(X=x) = \frac{3!}{x!\,(3-x)!} \pi^x (1-\pi)^{3-x}$$

$x = 0, 1, 2, 3$

$$\boxed{\therefore \text{prop}(X=x) = \frac{n!}{x!\,(n-x)!} \pi^x (1-\pi)^{n-x}}$$

(Table II)

$x = 0, 1, 2, \ldots, n$ = # of 1's out of n

This is the mass function, $p(x)$, of a binomial variable X.

E.g. X = # of heads out of n tosses

# of girls in a sample of size n.

Because we derived the above expression using proportions, it follows that $\sum_x \text{prop}(x) = 1$.

What do coins have anything to do with stats?

Recall, how we "derived" the binomial dist:
... by thinking about tossing a coin n times (or tossing n coins one time) and counting the number of heads, x. The statistical analog is taking a sample of size n, and counting the number of girls. $p(x)$ gives the proportion of times, out of some large number of repeats, that we would get x heads out of n tosses, or x girls in a sample of size n.

What's π?

For the coin example, it's the prob. of getting a H on one toss.
In the other example, it's the prob of drawing a girl, but that's also equal to the proportion of girls in the pop.

Don't confuse the various proportions:
- $p(X = x)$ ← Important
- π ←
- prop. of 1's in each sample of size n

This prop. is irrelevant in the binomial dist. But it will show-up later (Ch. 7, 8).

hw-lect 7-1

Make boxplots for each of the 2 continuous variables in hw-lect1. Compare and interpret the results, as we did in class.

Do it by R: boxplot(x, y) will make a figure with 2 box plots side-by-side.