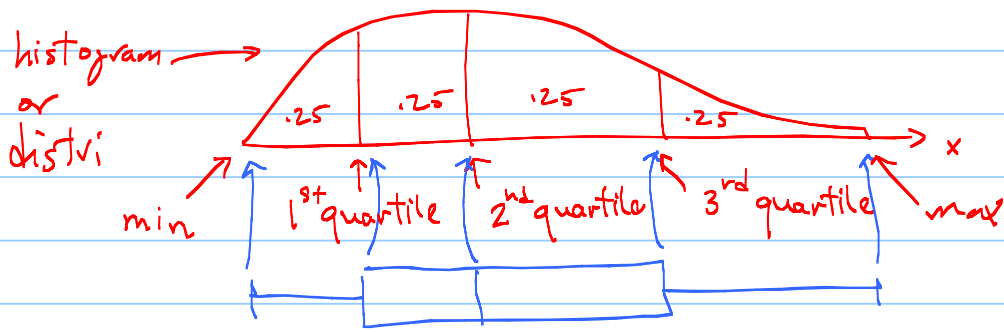
Lecture 6 (Ch. 1 mostly)

Last time we introduced the concept of the n^{th} percentile (for the normal distr.); an x value with $n\%$ area to its left. Note that percentiles (or quantiles, quartiles, ...) apply to dists and hists.

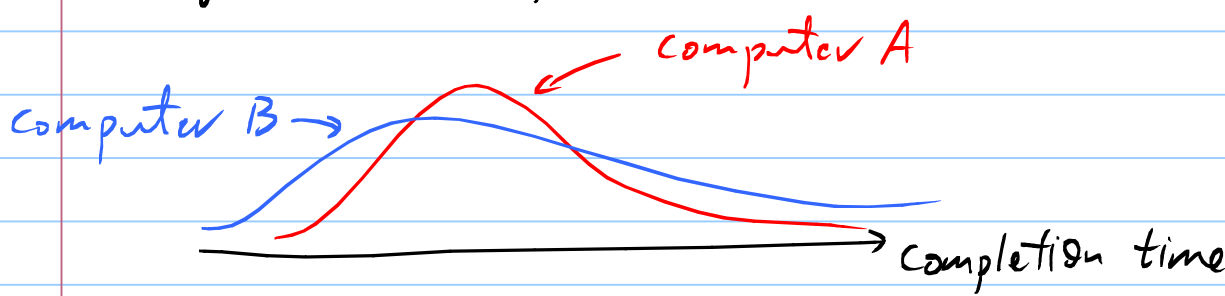
Quartiles are the basis of of the so-called "5-number summary" of a hist (or dist), often plotted as a boxplot:



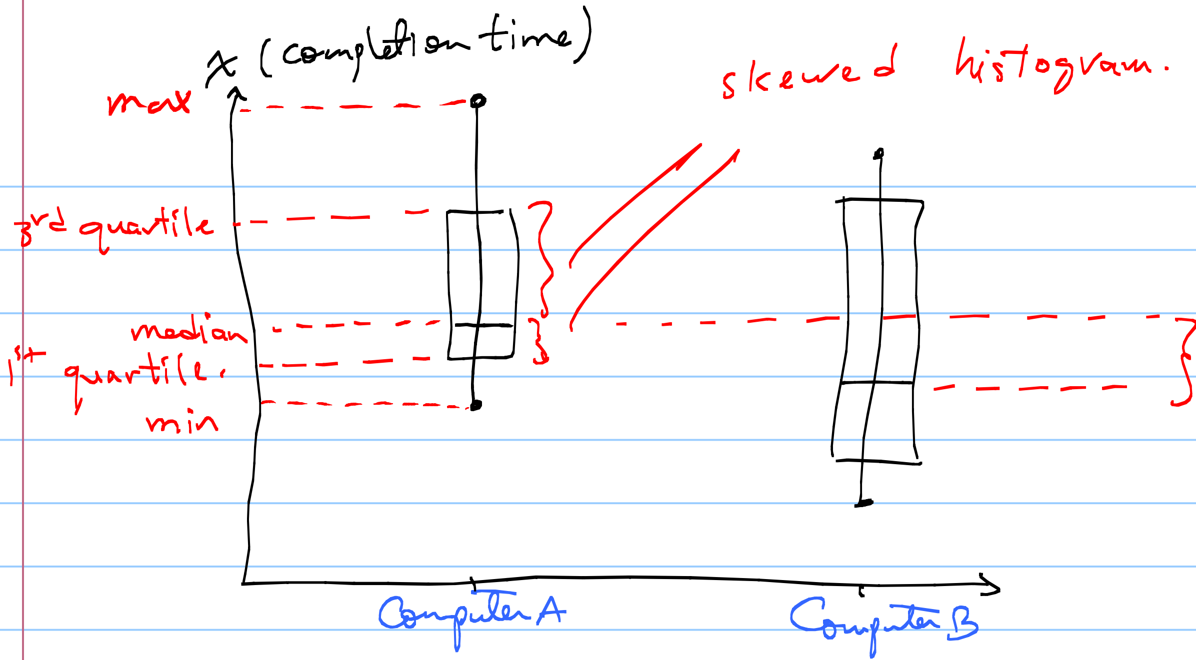
This material is from 2.3, but fits better here for lect. & Lab.

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

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 times 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 even involving many more (than 2 or 3) computers.

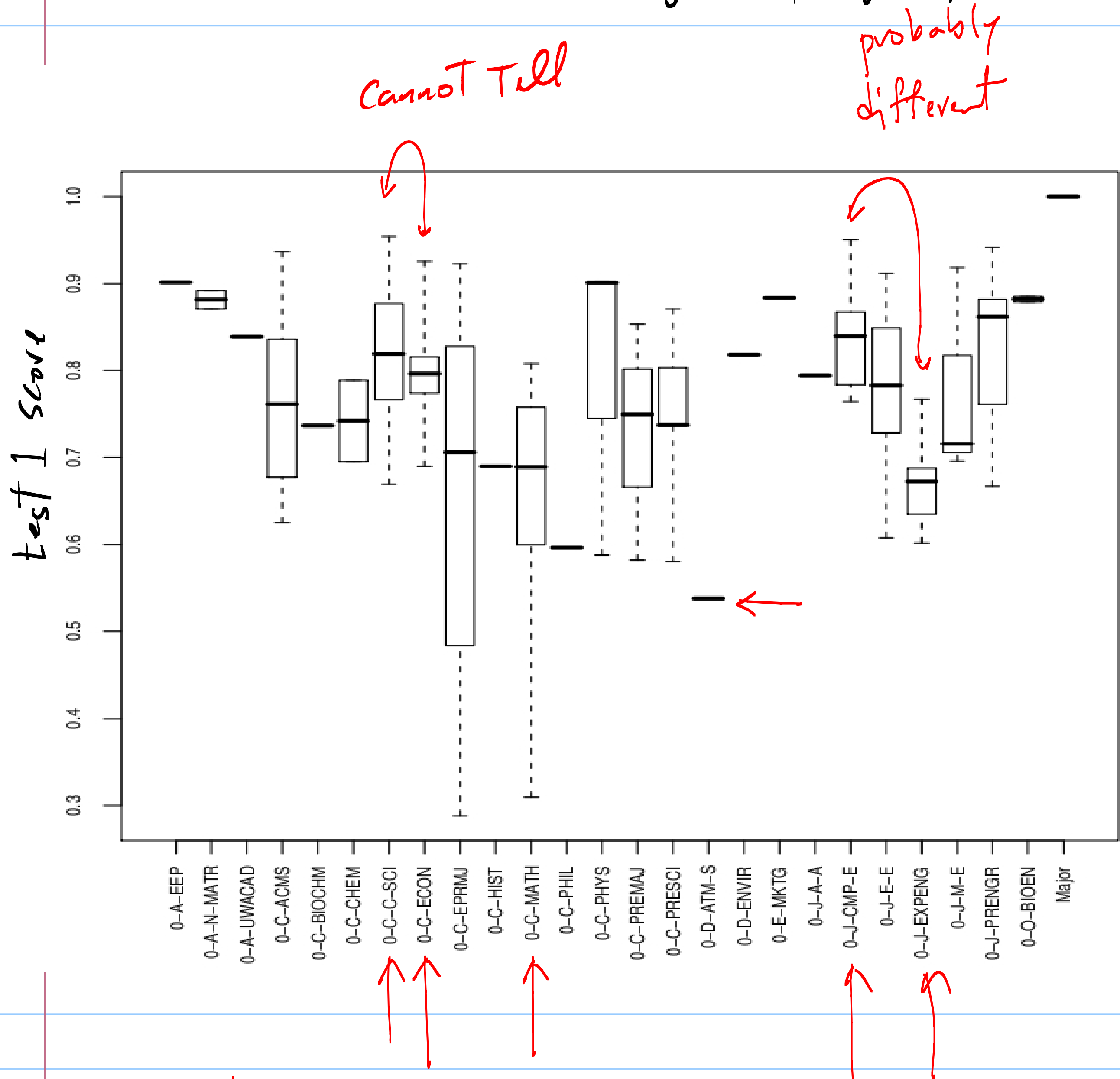


Observations: Based on This sample, computer B is faster "on average" because its median completion time is shorter. But computer B is also more "moody" (less consistent), because it has a wider spread in completion times. Important: Note spread !!

Having said all that, one cannot conclude that computer B is faster, because these boxplots are based on a sample of size 100. We do not 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 cannot conclude the B is faster. We cannot conclude anything! 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 :

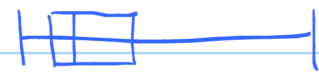


class discussion!

Note That The discussion involves comparing The whole boxplots, not comparison of The 5 numbers one by one.

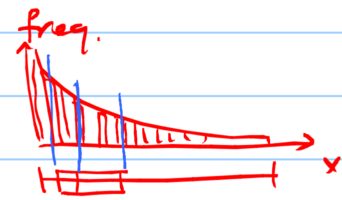
In summary, (comparative) box plots form a powerful tool of visually comparing multiple groups in terms of either data/sample from each group or Their distributions.

Here is a clicker question from the past:

Consider this boxplot of a histogram 

The histogram itself will "look"

A) Uniform   B) Normal $\boxed{\text{C) Exponential}}$ D) cannot tell



In the above question, one can also conclude that the population (ie. distribution) from which the sample was drawn may be 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. Next.

Derivation of 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 <u>not</u> asking for the prop. of 1's in <u>each</u> sample.

I'm asking for the prop., <u>out of the 10^8 trials</u>, ← i.e. large!

that are 1,1,1. Etc

			$X = \#$ of 1's
A	prop. of 1,1,1	$\overset{\text{independence}}{=} \pi \cdot \pi \cdot \pi$	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
	Etc.		
	0,0,0	$= (1-\pi)(1-\pi)(1-\pi)$	0

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

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

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

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

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

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

$x = 0, 1, 2, 3$

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

$x = 0, 1, 2, \cdots, n$ = # 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

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

Recall the connection between coin tosses and sampling:

The prob. of getting x heads out of n tosses of a coin (or 1 toss of n coins)

$\Updownarrow$

The prob. of getting x boys out of a sample of size n.
" " " x defective gates on a chip with n gates
Etc.

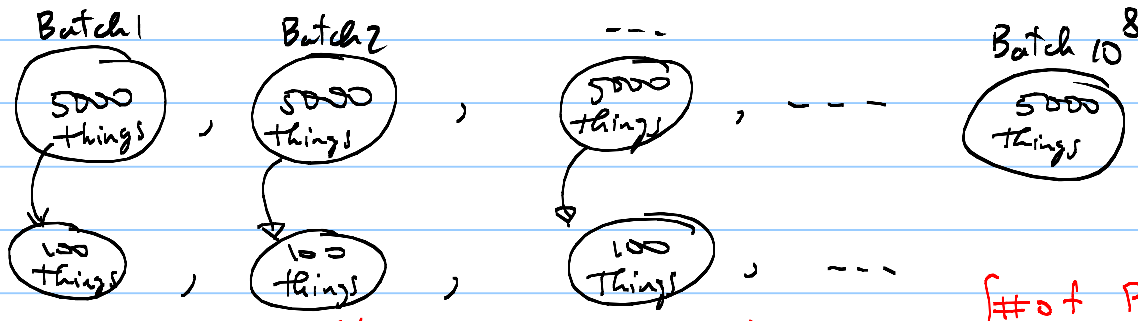
What's π?

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 boy,
i.e. the proportion of boys in the pop.

Don't confuse the various proportions: $\left\{ \begin{array}{l} p(X=x) \\ \pi \\ \text{prop. of 1's in each sample of size } n \end{array} \right.$ Important!

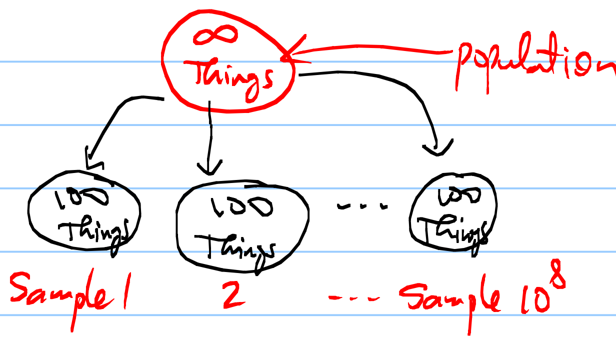
Irrelevant! It does not show-up in Binomial. It will later (Ch 7, 8).

Example 1.23 (p.56)



X = 3 (e.g.) , 5 (e.g.) , 0 (e.g.) --- = {# of Bads in sample of size 100.

Assume The lots are identical, ie.
The company manufacturing The 5000
Things is extremely consistent.
Then, The picture looks like This:



Q What proportion of These 10^8 lots will have $X = 0, 1, ---, 100$?

G = Good
B = Bad

Sample = $\{\underbrace{G, G, ---, G}_{100}\}$

Sample = $\{\underbrace{B, B, ---, B}_{100}\}$

Suppose we know The prop. of Bads, period, in The pop. = 0.5% = .005 = π

Then $P(X = x) = \binom{100}{x} \pi^x (1-\pi)^{100-x}$

prop. of lots with $X = 0$: $\binom{100}{0} \pi^0 (1-\pi)^{100}$ = .6058

$= 1$: $\binom{100}{1} \pi^1 (1-\pi)^{100-1}$ = .3044

$= 2$: = .0757

$= 3$: Etc. = .0124

$\vdots$

Lab.

Important Interpretation

In The long-run we expect
- ~ 60% of The lots to be all good.
- ~ 30% " " " to have 1 bad out of 100.
- ~ 7% " " " 2 bads " " ".
  (ie. 7% of The lots to be 2% defective)

hw-lect6-1 By R

Consider one of The two continuous variables, and one of The two discrete variables, in hw-lect 1. Make comparative boxplots for The continuous variable for each level of The discrete variable. E.g. if The discrete var. has 4 levels, Then you need to show 4 boxplots for The cont. var. all on The same plot, side-by-side. Interpret!

hw-lect6-2 By R

Today, in class, a student (who we shall call Chris) asked about The effect of sample size on boxplots. I said That boxplots are generally <u>not</u> affected by sample size. Let's see That. Write code to

a) take a sample of size 20 from a normal with $\mu = 0$, $\sigma = 10$.

b) " " " 30 "

c) 40

d) " 50 "

e) 100

f) Make a comparative boxplot of The 5 samples a-e.

hw-lect6-3

a) Use The binomial mass function to show That The prob. of getting "at least 1 head out of n tosses" is $1-(1-\pi)^n$, where π is The prob. of getting a head on a single toss.

Show work!

b) What is The numerical value of That prob. as $n \to \infty$? Think about The answer you get; it's interesting and counterintuitive.