Lecture 6 (Ch. 1, 2)

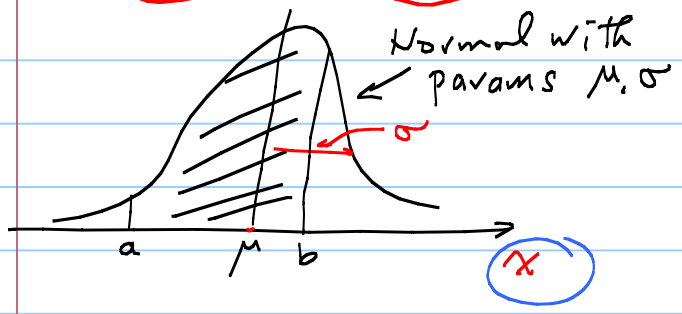
Last time we learned how to find areas under $N(0,1)$. "probs" from 2 left-areas Table I.

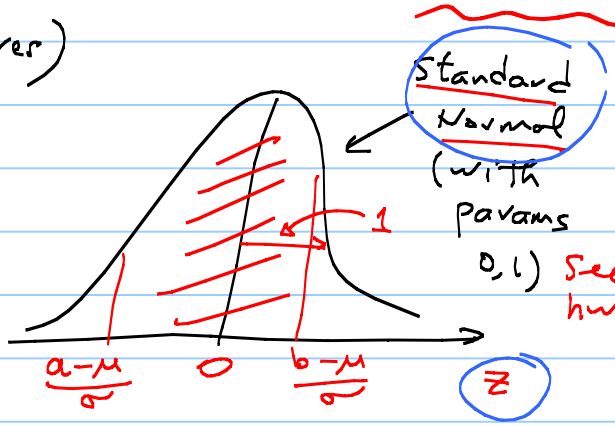
What about $N(\mu, \sigma)$?

It would be impractical to have tables for all μ and σ!

Fortunately, there is a trick: Change variables! also called standardization

$$x \to z = \frac{x-\mu}{\sigma} \quad (z\text{-score})$$



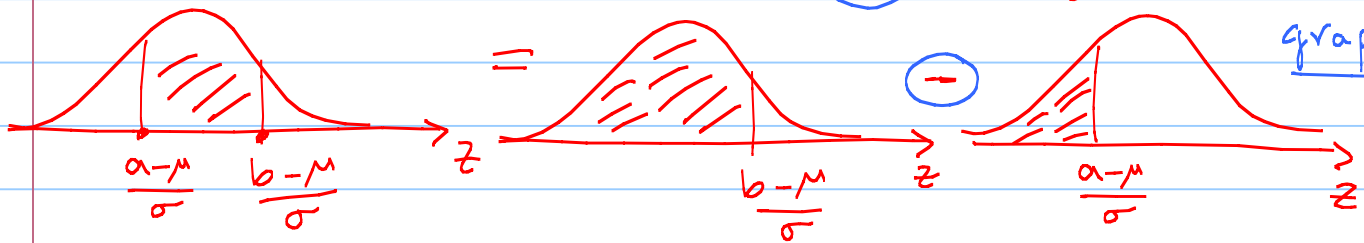


So, to compute area between 2 values:

Important first step: Standardize!

$$\text{prob}(a < x < b) = \text{prob}\left(\frac{a-\mu}{\sigma} < \underbrace{\frac{x-\mu}{\sigma}}_{z} < \frac{b-\mu}{\sigma}\right)$$

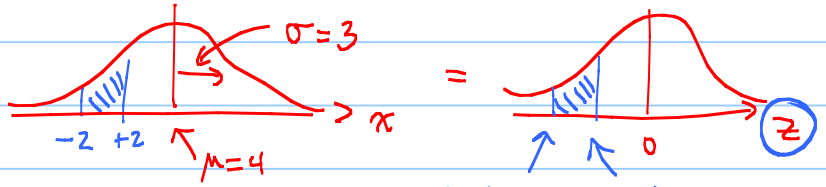
graphically



$$= \text{prob}\left(z < \frac{b-\mu}{\sigma}\right) - \text{prob}\left(z < \frac{a-\mu}{\sigma}\right)$$ Algebraically

Either way (algebraically or graphically) you can obtain the value of each term from Table 1.

Example: If $x \sim N(\mu=4, \sigma=3)$, what's $pr(-2 < x < 2)$?



$$-2 = \frac{-2-4}{3} \qquad \frac{2-4}{3} = -0.67$$

$$= 0.2514 - 0.0228$$

$$= \boxed{0.2286}$$

Standardizing with $x \to z = \frac{x-\mu}{\sigma}$ is specific to $N(\mu,\sigma)$:

$$x \sim N(\mu,\sigma) \longrightarrow z \sim N(0,1)$$

The main point is that we can then find $pr(\ a<x<\ )$: (See hw)

$$pr(\ a<x<b) = pr\left(\frac{a-\mu}{\sigma} < z < \frac{b-\mu}{\sigma}\right).$$

But, more generally, any change of variable, $x \to y$, that allows you to compute $pr(\cdots<x<\cdots)$ is useful (and can be called "standardizing").

distribution percentile (or quantile)

So far: Find area, given x

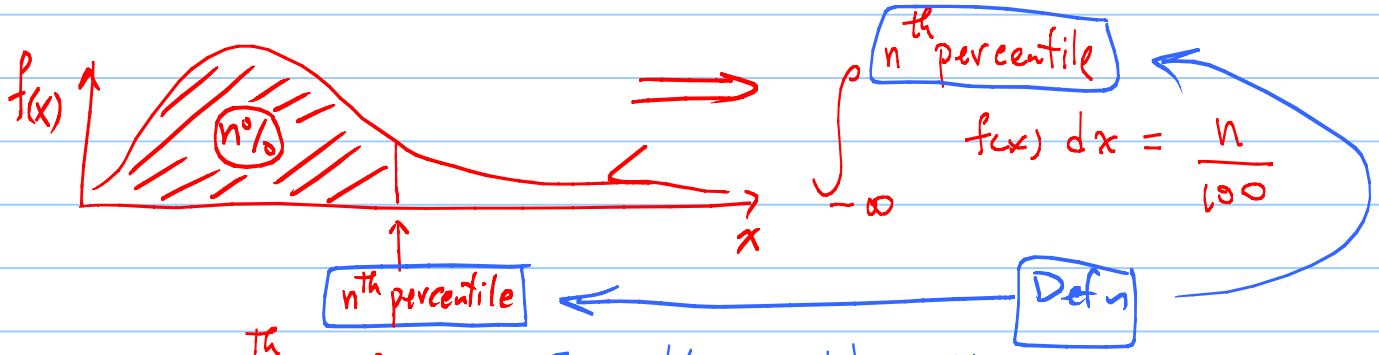
Now: Find x, given (some left/right) area.

The simplest example is the distribution median:

area = 0.5 = 50%

E.g. median: $\int_{-\infty}^{\text{median}} f(x)\,dx = \frac{50}{100}$

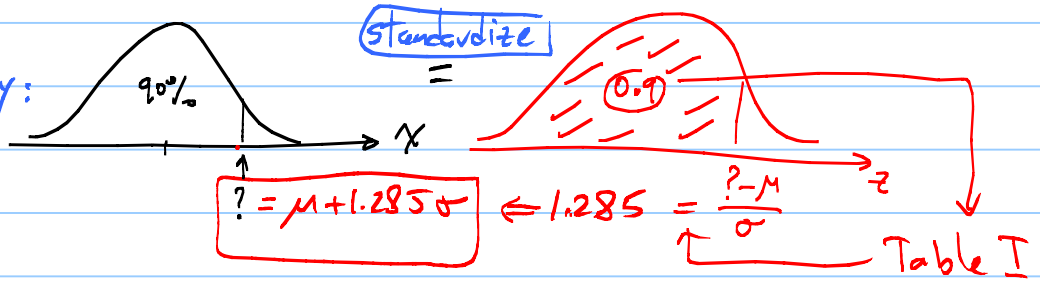
Median is a special case of the more general concept of percentile:



median = 50th percentile = 0.5 quantile = 2nd quartile

Example: What's the 90th percentile (0.9 quantile) of $N(\mu,\sigma)$?

Graphically:



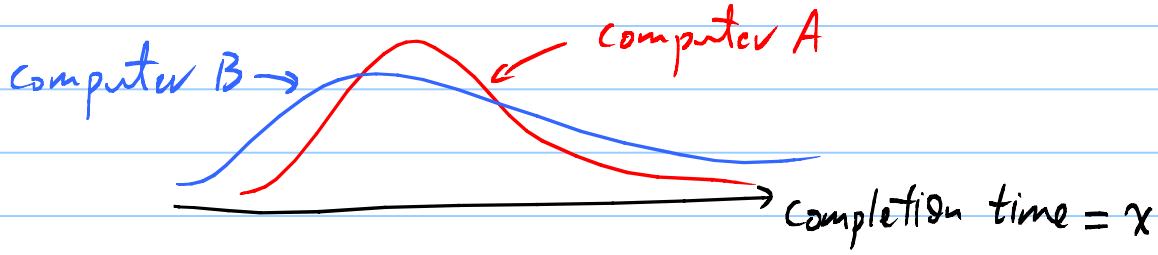
Algebraically: $0.9 = pr(x < ?) \overset{\text{Standardize}}{=} pr\left(\frac{x-\mu}{\sigma} < \frac{?-\mu}{\sigma}\right) = pr\left(z < \frac{?-\mu}{\sigma}\right)$

Table I $\rightarrow \frac{?-\mu}{\sigma} = 1.285$

$? = \mu + 1.285\sigma$

Important comments:
1) The nth percentile/quantile/quartile/... is a number on the x-axis in the above figure. In other words, it has the same units as x itself. If x is in Kg, then the nth percentile is a number in Kg. It is NOT (necessarily) a percent.
2) The notion of the nth percentile/quantile/quartile/... applies to histograms, as well. In that case, it's called the sample percentile/quantile/quartile/... . So, for example, the 65th sample percentile of data on weight is a specific value of weight in your sample for which 65% of the cases are smaller.
3) The notion of the nth percentile/quantile/quartile is an extremely useful concept. It shows-up everywhere. One place is in summarizing distributions and histograms. Look:

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/should then look at the histogram of "completion time" for the 2 computers!



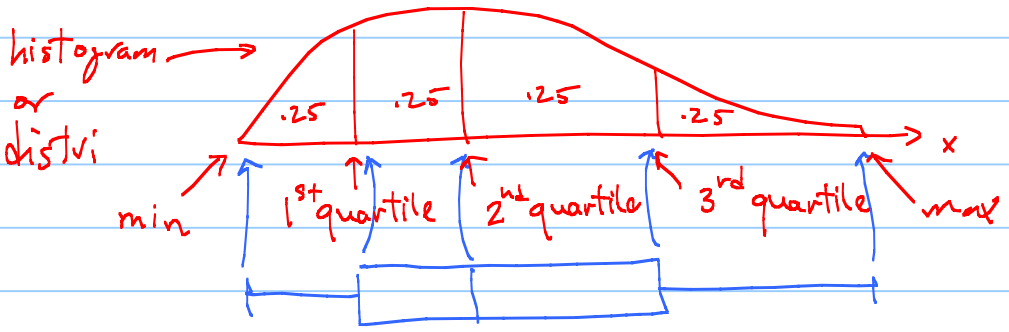
Which computer is faster? Discussion:

1) It looks like B is faster on the average. But B is also more moody! (Learn to also look at the width of histograms.) So, which computer should you buy?! The answer: It depends.

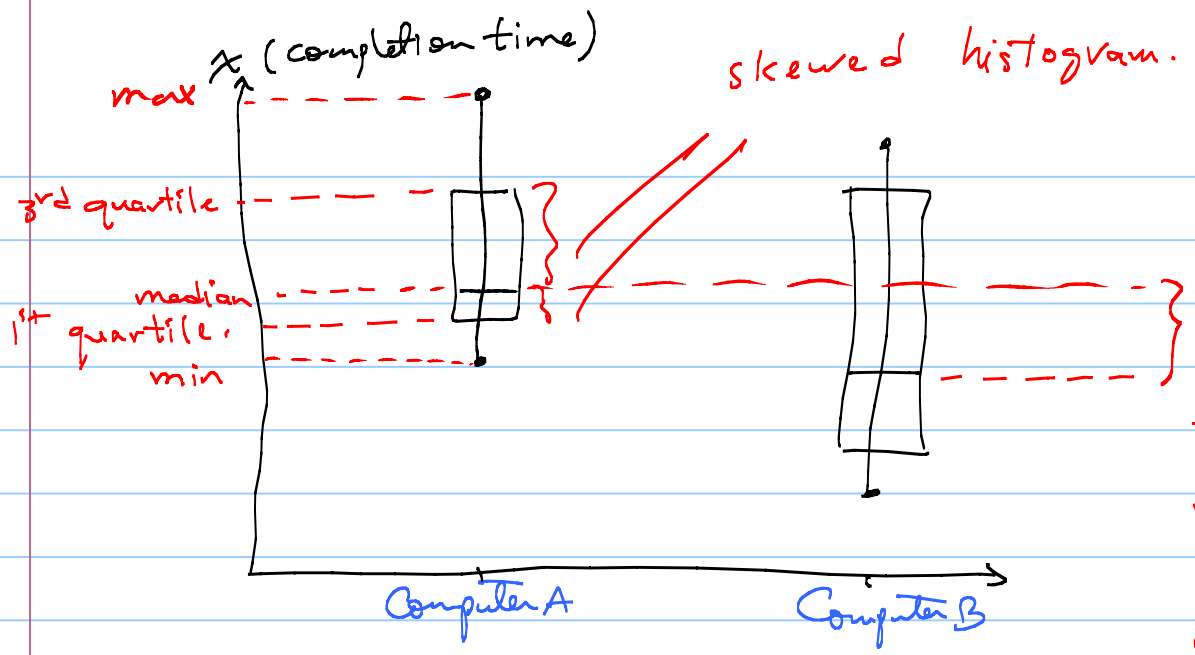
2) More importantly, the huge overlap between the histograms causes a huge problem. The true mean of x for the two computers is somewhere close to the center of each histogram; but we don't know where After all, our data is only a sample taken from some unseen population. So, if/when there is too much overlap, then we cannot tell which computer is truly faster; see the next page, too.

Comparison of even 2 groups is complex, involving location, width, ...
How do we handle more than 2 groups?

Quartiles are the basis 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.



Observations: Based on this sample, we cannot really tell if one computer is faster. We could have been able to say something if one hist/boxplot were strictly shifted with respect to the other. Note the the REASON we caanot tell is the WIDTH of data/hist/boxplot. This is one way in which the width of data plays an important role.

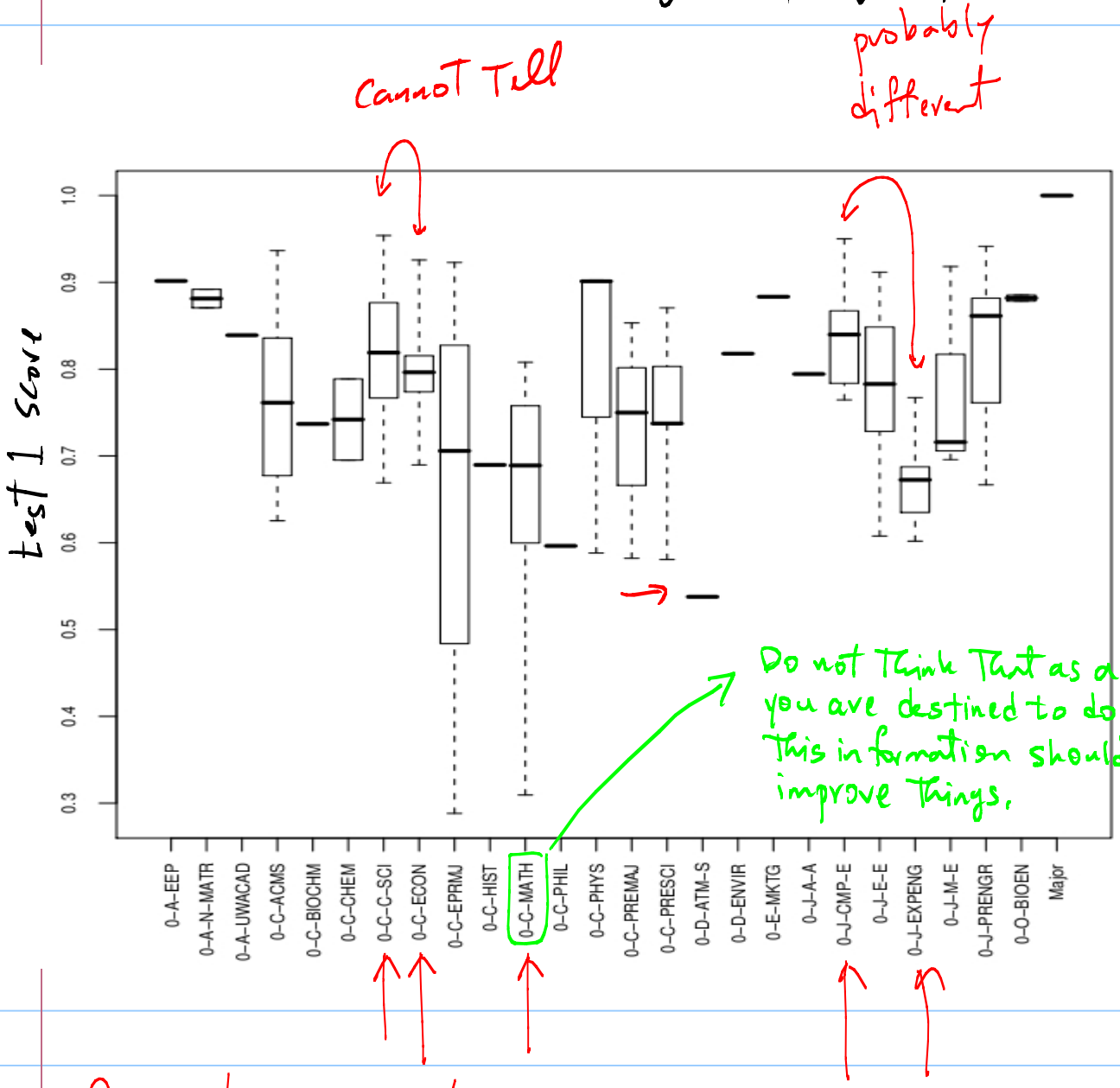
Now. suppose we decide to compare the computers only in terms of the center (say, median or mean). Then, computer B is faster "on average" because its typical completion time is shorter. But computer B is also more "moody" (less consistent), because it has a wider spread in completion times. So, again, WIDTH plays an important role.

Having said all that, one cannot conclude that computer B is faster, because these boxplots are based on a sample/ hist. We do not know what is the distribution of x (i.e. the population). We do know the true dist./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 that B is faster in the population. In fact, in this case we cannot conclude anything about the population/distribution because there is too much overlap. Get used to this kind of conclusion! It happens because of the existence of WIDTH!

How much overlap is too much? Ans. in Ch-7 and beyond. For now, just learn that every time 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/width.

READ!

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) boxplots form a powerful tool of visually comparing multiple groups in terms of either data/sample from each group or Their distributions.

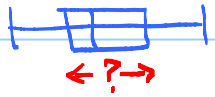
hw_lect6-1 If x follows $N(\mu,\sigma)$, what's the prob. of x being within 1 σ of μ? Get used to this jargon.

hw_lect6-2 Standardization is important in finding probs. Although it almost always refers to the change of variable $z=\frac{x-\mu}{\sigma}$, taking $N(\mu,\sigma)$ to $N(0,1)$, sometimes a different change of variable is required to obtain something that has $N(0,1)$ distr. Find the $pr(x<2)$ if $\left(\frac{1}{x}\right)$ has a Std. Normal dist.

hw_lect6-3 Another of the named distributions is the so-called power-law dist. It's formula is $f(x)=\alpha x^{\alpha-1}$, $0<x<1$, where $\alpha>0$ is its parameter.

a) Find the n^{th} percentile of that distribution. Hint: The answer will depend on α and n.

b) How long is the box portion of the corresponding boxplot? Hint: The answer will depend on α.



hw_lect6-4

Consider ONE of the 2 continuous random vars, and ONE of the 2 discrete/categorical variables in the data you collected. 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 continuous variable, all on the same plot, side-by-side. Interpret/discuss them. By R.