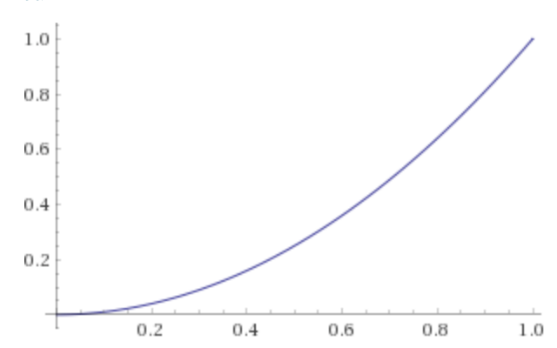
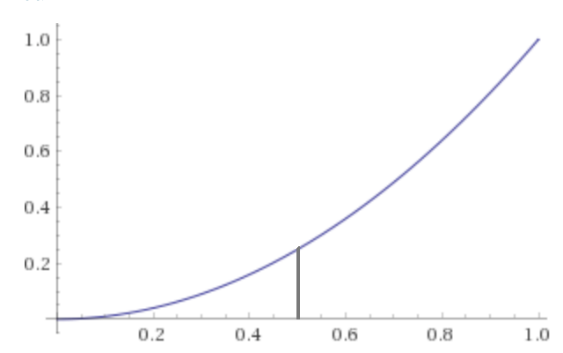
**Part 1: Preliminaries**

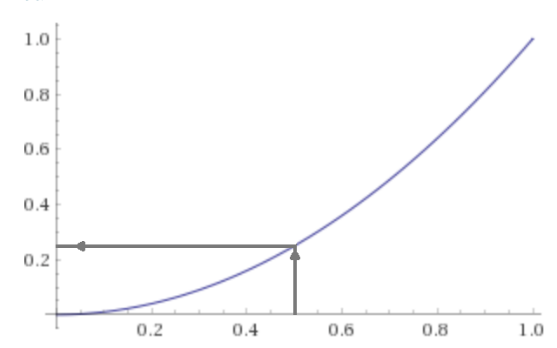


This... is a function!  (It gets more interesting from here, promise.)  Or more precisely, it's a graph of a function.  You render a function F's graph by putting in a bunch of x values and drawing the point (x, F(x)) for each one.  So in this case, for x^2, you get values like (0,0), (0.1, 0.01), (0.2, 0.04), and so on, all the way up to (1,1).

Among other things, this means that if we want to "apply" the function we've just graphed, we can pick a point on the x axis and follow it up to the curve:

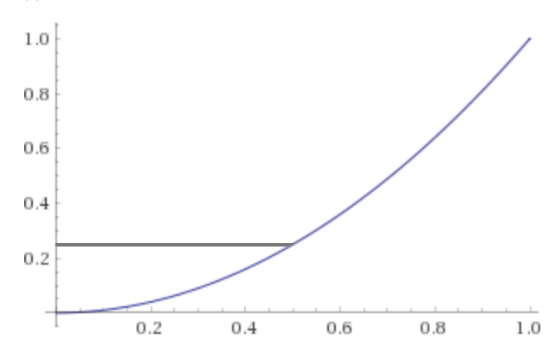


...and figure out where that point falls on the y axis:

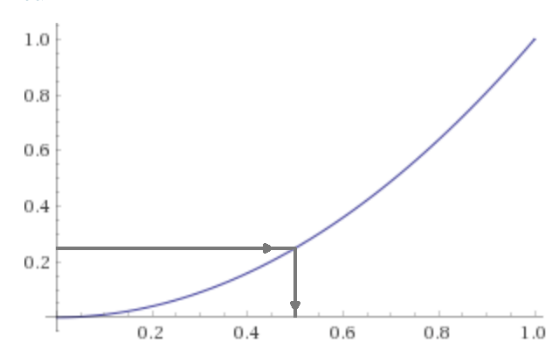


Voila!  F(x).  In this case, F(0.5) = 0.25.

Now, here's the less-obvious corollary: if we want to find F^[-1](y), we can do something remarkably similar.  Pick some point on the y axis and follow it over to reach the curve:



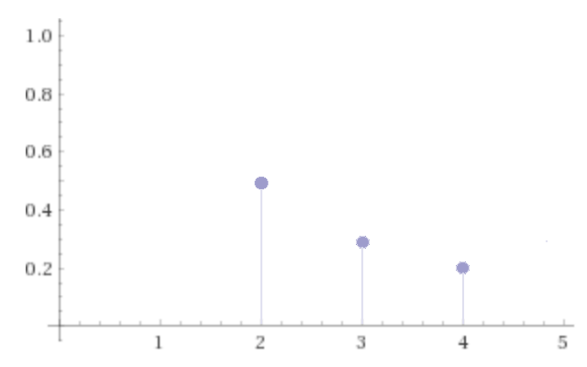
Then figure out where that point falls on the x axis:



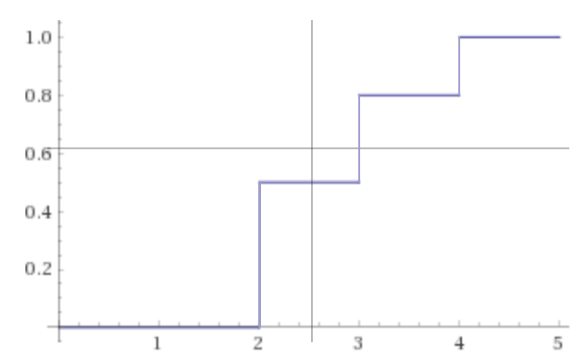
Voila!  F^[-1](y).  In this case, F^[-1](0.25) = 0.5.

**Discrete distributions**

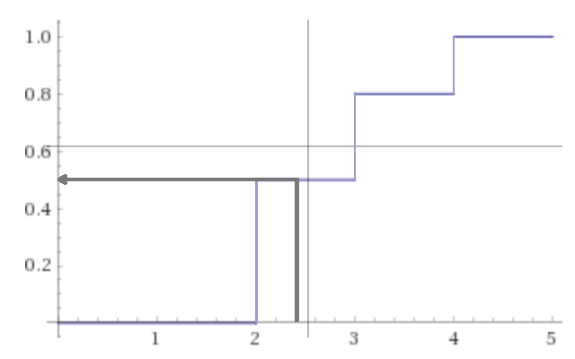
We like to describe [discrete] probability distributions with one of two functions, depending on how we're using them.  One function, the pmf f(x), captures the probability of sampling any given value from the distribution.  So for instance, if I tell you "I'm thinking of a number.  There's a 50% chance that it's 2, a 30% chance that it's 3, and a 20% chance that it's 4.", I would describe that with this pmf:



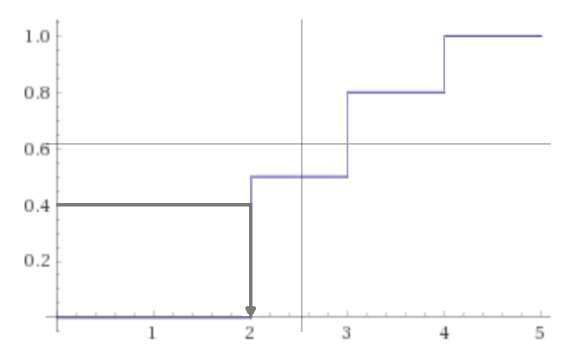
The other function, the cdf or *cumulative* distribution function F(x), describes the probability that any given sample from this distribution will be less than some threshold, x.  So for the distribution I described above, the probability of sampling a value less than 2 is 0.  The probability of sampling x < 3 is 0.5; the probability of sampling x < 4 is 0.5+0.3 = 0.8; the probability of sampling x less than, say, 5, is 1.  (This makes sense, because all of the possible x values - 2, 3, and 4 - qualify, right?)



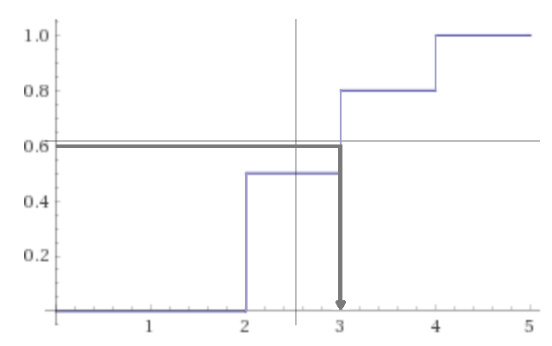
That's a plot of the CDF.. which is a function.  In part 1 we were looking at F(x) and F^[-1](y) using a plot.  No reason we can't do the same here!



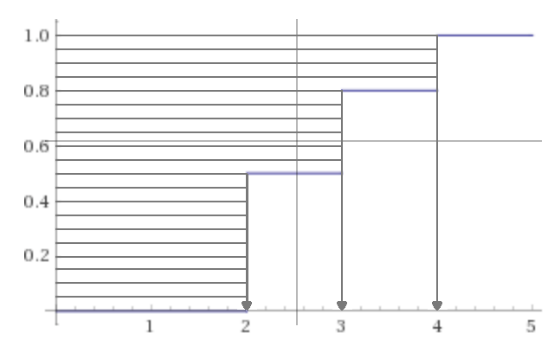
F(2.4) = 0.5, which checks out - for this distribution with possible values 2, 3 and 4, the probability P(x ≤ 2.4) is simply p(x = 2).1



F^[-1](0.4) = 2, which also checks out - F(2) > 0.4.  What happens if we do another one?



F^[-1](0.6) = 3.  Again, makes sense since F(2) < 0.6 ≤ F(0.3).  What happens if we do a whole bunch?



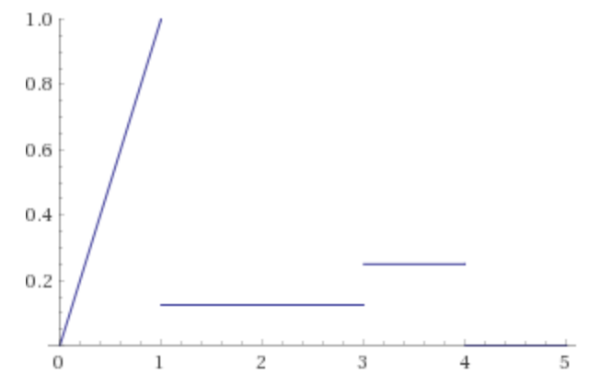
Well, a lot of them "land" at 2, some "land" at 3, and a few "land" at 4.

In fact, we can count exactly how many: 10 of 20 samples (or 50%), went to 2; 6 (30%) went to 3; and 4 (20%) went to 4.  For these uniformly distributed values in [0, 1], the proportion that "lands" on each number (when we take it through F^[-1]) is exactly equal to the change in height of F^[-1] at that number - which is itself equal to the probability mass associated with that number in the original f(x)!

**Continuous distributions**

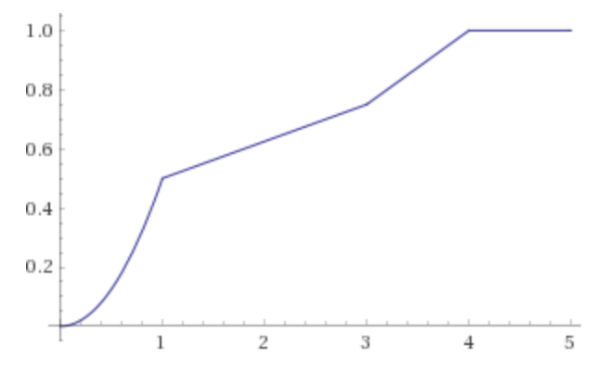
Continuous distributions can also be described with two functions.  One function, the pdf f(x), describes how relatively likely (or unlikely) we are to get various numbers by sampling the distribution.  The other function, the CDF F(x), works exactly the same way as for a discrete distribution - this function answers the question, "What is the probability that a sample from this distribution is less than (x)?"

Let's consider a completely arbitrary2 continuous distribution with the following pdf:



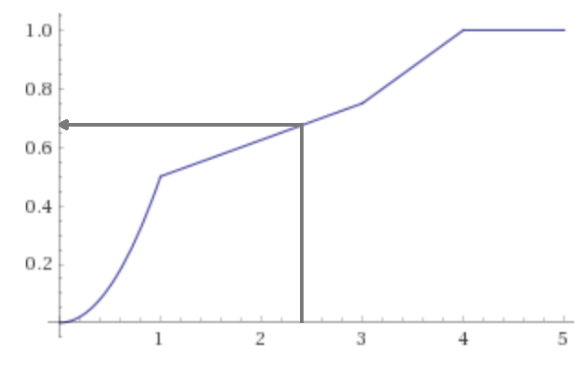
This distribution has some high-ish probability of sampling values between 0 and 1, which within that interval is very high near 1 and drops precipitously towards 0; a fairly low probability of sampling a given number between 1 and 3; a somewhat higher chance of sampling a given number between 3 and 4; and a 0% chance of sampling anything greater than 4 (or less than 0, but that part of the graph isn't shown).

This distribution can be integrated, and luckily already has been integrated, to give the following CDF F(x):



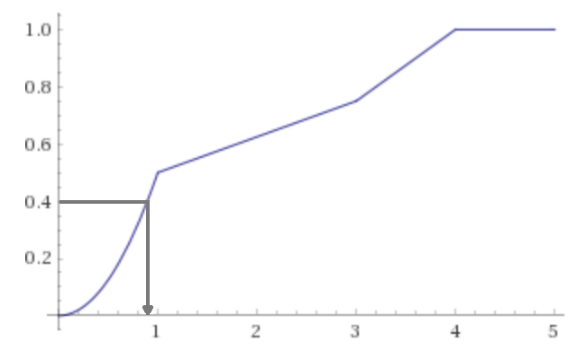
Ain't it a beaut?

Now, we can do all the same things with this CDF plot that we did with the other one; after all, they are both CDFs.  For example, let's approximate the value of F(2.4):

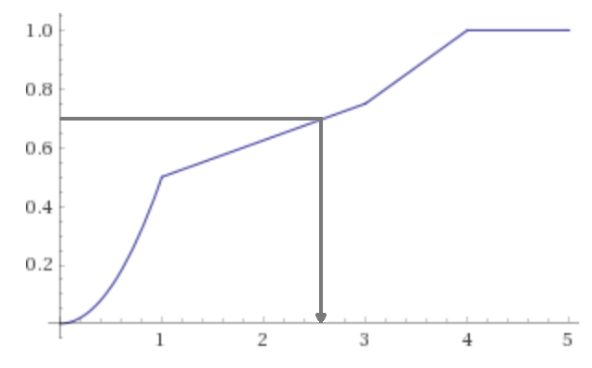


Looks like it's a little less than 0.7.  (In fact, it's exactly 0.5+0.25\*(1.4/2) = 0.675.3)

Or, we can approximate the value of F^[-1](0.4):

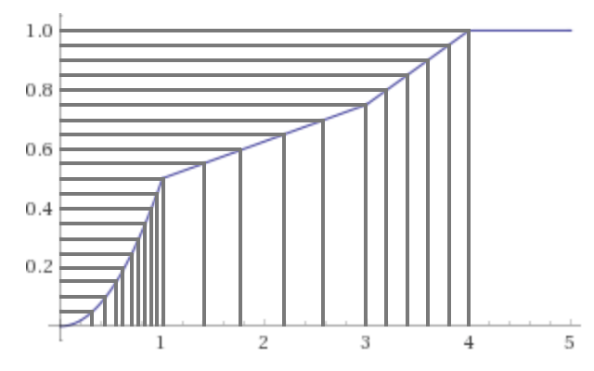


A little less than 1 - which again checks out.  (In fact, it's approximately sqrt(0.8), or ~= 0.89.3)  Here's another one:



F^[-1](0.7) = looks like about 2.6.  (In fact, it's exactly 2.6 - the formula is 1+(3-1)\*(0.7-0.5)/0.25.3)

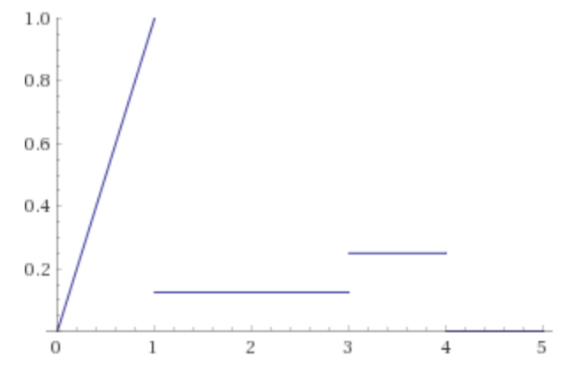
So what happens if we repeat that little exercise we did above, of taking a bunch of samples from the unit interval and mapping them back through F^[-1](y)?



Pretty!4  But what does it mean?

Well, you'll notice some patterns.  A lot of F^[-1](y) values end up in the range from 0 to 1, with especially many ending up close to 1.  There are relatively fewer that land in the 1-3 range, and the same absolute number land in the 3-4 range even though it's half the length in the x direction.

In fact, if we squint, the statement we made above when looking at the discrete CDF graph still seems to hold.  It doesn't really make sense to talk about the change in height at a particular point on the graph, but we can talk about the change in height over an interval - and the number of samples that 'hit' it is proportional to that change.  If taking a huge interval like [0, 1] or [3, 4] doesn't appeal, we can talk about the slope at a point on the graph - if we could somehow get the slope of every point, we could describe differences like "3.4 is twice as likely to be 'hit' as 2.4."  And by slope I mean derivative.  And by somehow, I mean by recycling a graph we already saw above, because...



The derivative of the CDF is the pdf.

**Bringing it all together**

So both types of distributions have a CDF F(x) and a pmf or pdf f(x).  The pdf describes the relative likelihoods of obtaining different values when you sample this distribution - that is, the likelihood of sampling value x from the distribution is proportional to f(x).

If we uniformly take a bunch of samples from the interval [0, 1] and send them through the F^[-1](y) looking glass, they will land on the X axis with a density described by f(x).  The more samples we take, the closer the final distribution will be to f(x).

If we take samples uniformly at random from [0, 1] - that is, if we sample U(0, 1) - and dump them into F^[-1](y), what happens?  Well, the likelihood of sampling value x will be proportional to f(x).  But this is exactly the definition of sampling a distribution with pdf or pmf f(x)!

Hence, the inverse transform theorem, which is handy for any distribution where you can obtain F^[-1](y) - regardless of its type (or ugliness).  Maybe continuous and discrete distributions aren't so different after all.5

-----

1Recall that a discrete distribution is one where you can list the possible outcomes in some order.  Among other things, this means that you can get the value of the CDF by (1) listing the possible outcomes, (2) figuring out which ones are relevant (i.e. less than your value x) and (3) summing their probabilities.

2So arbitrary, in fact, that I regretted the choice several times during the unexpectedly painful and time-consuming process of marking up the graphs.  Future distributions will be arbitrarily selected from {f(x) = 2x, f(x) = 3x^2}.

3The explanation is left as an exercise for the reader and/or Michael Kuehn.

4(cute little arrows omitted because they turned this thing into an illegible mess)