Phylomath Lecture 4

Brigid O'Donnell (17 February 2004)

A return to sojourn times

We began by returning to the idea of sojourn times, or the time period until the next disruption event occurs for a given site in a nucleotide sequence. We're interested in understanding sojourn times in order to ultimately calculate the number of disruptions (or substitutions) that occur over a single branch in a phylogeny. For a binomial distribution, the expected number of substitutions (signified by λ) on that branch equals the product of the 1) probability of a substitution in the interval, *p*, and 2) the number of intervals the branch is divided into, *n*:

$$\lambda = pn$$

For a Poisson distribution, the expected number of substitutions (signified by λ) for a specified branch equals the product of 1) the instantaneous rate of disruption, denoted μ, and 2) time, t:

$$\lambda = \mu t$$

But how are these two λ values related to each other? Or, how do we compare these two values to each other when it seems they are using terms that are not directly comparable? We can do this by first letting the number of intervals (*n*) in the binomial distribution approach infinity over a very small time period (we'll call this "*dt*") such that now our *t* value can be seen to equal

$$t = (dt)(n)$$

Substituting this new definition of *t* into the Poisson version of λ leads to the logical conclusion that:

$$\mu t = \mu(dt)(n)$$

The product of the first two terms, μ and *dt*, is proportional to *p*. This term equals the instantaneous probability of a disruption, or the expected number of disruptions in amount of time equal to *dt* (an incredibly small amount of time). The product of *dt* and *n* is proportional to time, *t*. We now can see how the two distributions are related.

Expected number of disruptions over a time period *dt*

If we want to figure out the expected number of disruptions occurring over a small portion of that branch, called *dt*, we would see that this expected number actually equals the probability of a disruption occurring in the first place. We'll show how this is true below.

Letting *y* equal the number of disruptions, and E(*y*) equal the expected number of disruptions, we can prove that the expected value is equal to the probability of a disruption. First, we stipulate that there cannot be greater than 1 disruption in the tiny interval of time, *dt*, so our possibilities are 0 disruptions and 1 disruption occurring in that tiny time period *dt*. Thus, in order to calculate the expected number of disruptions that occur, we calculate the product of the value and the probability of that value, for each possibility, and then sum up these values:

$$E(y) = (0)pr(0) + (1)pr(1)$$

$$E(y) = 0 + pr(1)$$

$$E(y) = pr(1)$$

Therefore, the expected number of disruptions equals the probability of a disruption occurring. We can predict what the expected number of disruptions is from information about the rate of a disruption occurring and the amount of time that transpires (i.e. $\lambda = \mu t$). This is analogous to the idea that in traveling someplace, we can predict the distance we will travel if we have information on the speed at which we are traveling and the time we will be in transit.

Note added by Paul: The expected number of disruptions over the tiniest of time intervals *dt* is μdt (rate multiplied by time). This section was about showing that this expected number of disruption equals the probability of a disruption when you are envisioning a time period so small that only 0 or 1 substitutions are possible. Hence, $p = \mu dt$ (where p harks back to the binomial model, where we let *p* be the probability of a disruption in one of the *n* arbitrary time intervals). The first two sections in these lecture notes are simply showing that you can arrive at the conclusion that $\mu\ dt$ is the instantaneous probability of a disruption in more than one ways. I am spending some time on this subject because it is important to remember that it is $\mu\ dt$ that is the instantaneous ***probability*** of a disruption, not just μ by itself, which is the instantaneous ***rate*** at which disruptions occur.

Probability density functions & Cumulative distribution functions

Paul next introduced the concepts of the probability density function and cumulative distribution function. The exponential probability density function (PDF) is proportional to the probability that a disruption occurred ***exactly*** at time *t*, ($T = t$), and can be shown as:

$$f(t) = \mu\ e^{-\mu t}$$

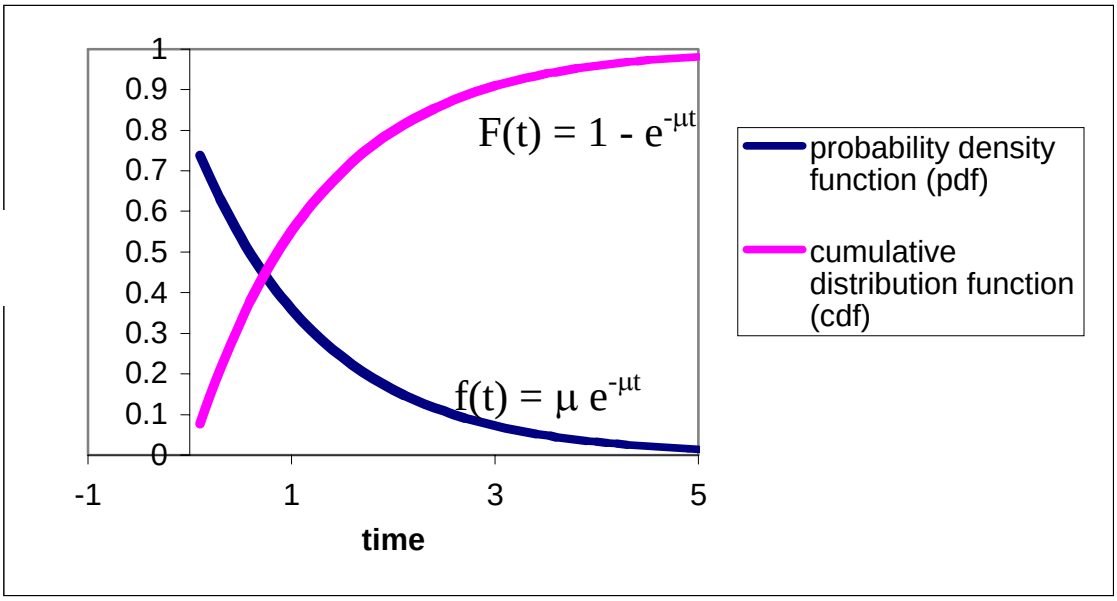
The exponential cumulative distribution function (CDF) is proportional to the probability of at least one disruption occurring from time 0 to time *t*, ($T \leq t$) and can be shown as:

$$F(t) = 1 - e^{-\mu t}$$

NOTE: *T* (in capital letters) here refers to a random variable (e.g. leaf length), and *t* (small letters) refers to a particular instance/value of that variable (e.g. 2.34cm). Also, the conventional use of small *f* and big *F* is important as this distinguishes a probability density function from a cumulative distribution function.

Note added by Paul: The CDF and PDF concepts are not restricted to disruptions or even biology. They are central concepts in statistics as a whole, whether Bayesian or frequentist. I was using sojourn times as examples of random variables, and sojourn times are exponentially distributed (as we discovered when we worked out the form of the CDF and PDF).

Let's visualize these functions graphically (note: $\mu = 0.8$ for these examples):



The y-axes for each of these functions are different; for the probability density function the value on the y-axis is the probability ***density***, but for the cumulative distribution function the y-axis value represents a ***probability***. Probability densities can be larger than 1 (although this is not true in this example), whereas probabilities must be between 0 and 1.

The **cumulative distribution** F(t) can be approximated as the sum of the areas of all the small rectangles fitting under this curve between 0 and t,

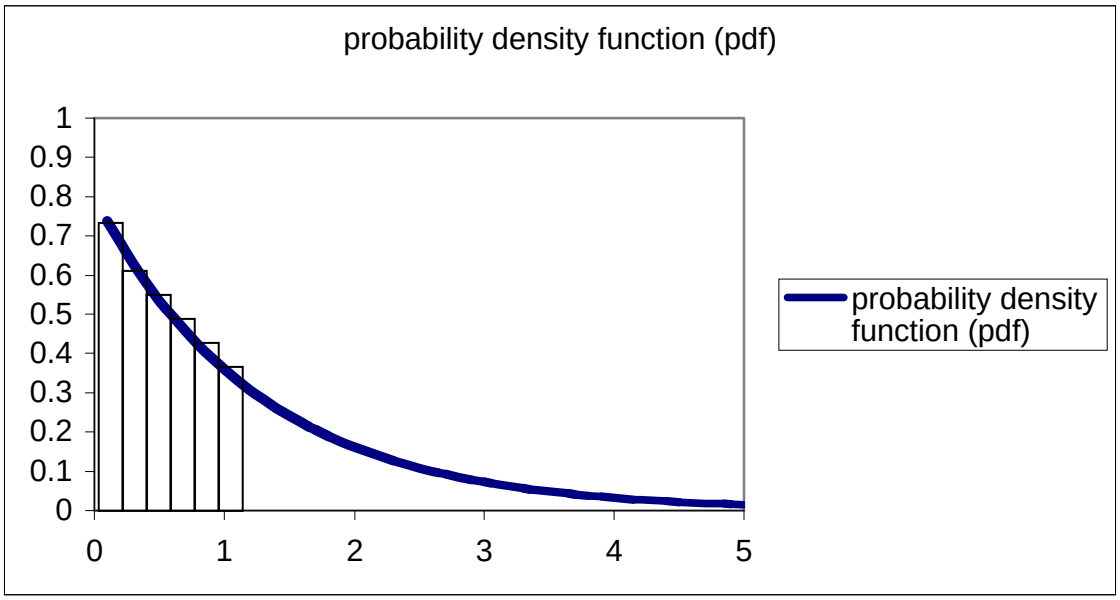
$$F(t) = \Sigma_i f(t_i)\, \Delta t,$$

where Δt is the width of each rectangle and $f(t_i)$ is the height of rectangle *i*.

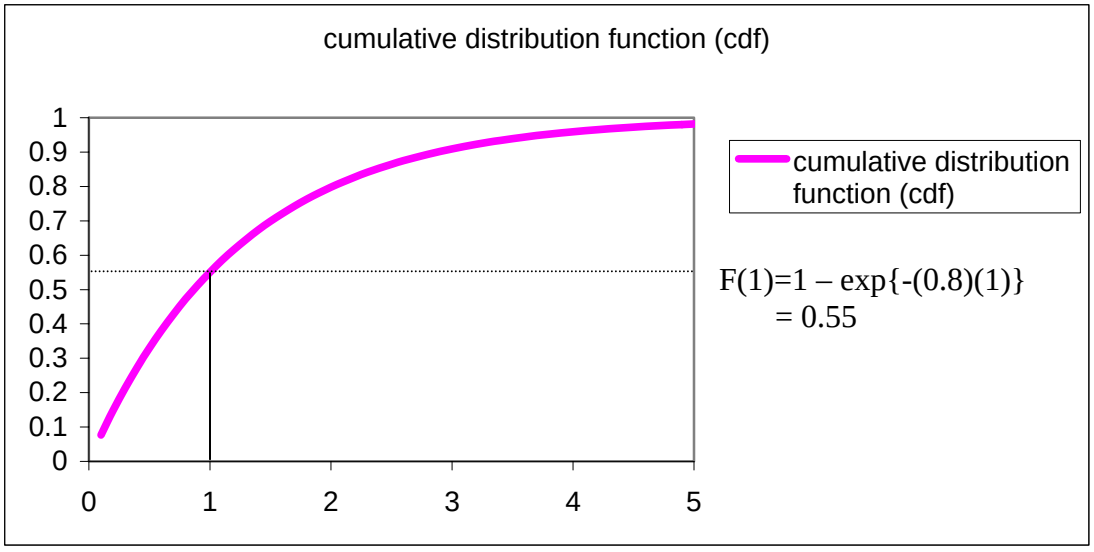
If Δt is allowed to approach zero, summing up all of the tiny rectangles from time 0 to time t becomes equivalent to taking the integral under that portion of the curve:

$$F(t) = \int_0^t f(t)dt$$

Now that tiniest of time intervals, the ***differential*** *dt*, takes the place of Δt and f(t) is the height of the rectangle of width *dt* positioned exactly at time *t*. Let's look at this graphically. If we were interested in determining the probability of a disruption occurring at any time between time 0 and time 1 (where 1 is simply an arbitrary time chosen for illustration) we could first calculate the sum of the areas of all of the small rectangles contained under the probability density function curve over that time period.



This value is equal to the value of the cumulative distribution function evaluated at time 1. This value is equal to the height of the CDF at time *t*=1, or F(1):



Relating the CDF and PDF

To get the cumulative distribution function F(t) when the probability density function is known, we need to simply integrate the probability density function from 0 to *t*, and

similarly, to get the probability density function f(t), we need to differentiate the cumulative distribution function at the point *t*:

$$F(t) = \int_0^t f(t)dt \text{ and,}$$

$$f(t) = F'(t)$$

Since we have already know the equations for these functions (above), we can prove that this relationship holds by taking the derivative of the cumulative distribution function to arrive at the probability density function, as follows:

$$F(t) = 1-e^{-\mu t}$$

$$f(t) = F'(t) = \mu e^{-\mu t}$$

Eight useful rules for calculating derivatives

Note the convention that f '(x) means "the derivative of f(x)"

1. $f(x) = a$ (where a is a constant)

   $f'(x) = 0$

2. $f(x) = x^b$

   $f'(x) = bx^{b-1}$

3. $f(x) = a\ g(x)$

   $f'(x) = a\ g'(x)$

4. $f(x) = \ln(x)$

   $f'(x) = 1/x$

5. $f(x) = \ln g(x)$

   $f(`x) = g'(x) / g(x)$

6. $f(x) = e^x$

   $f'(x) = e^x$

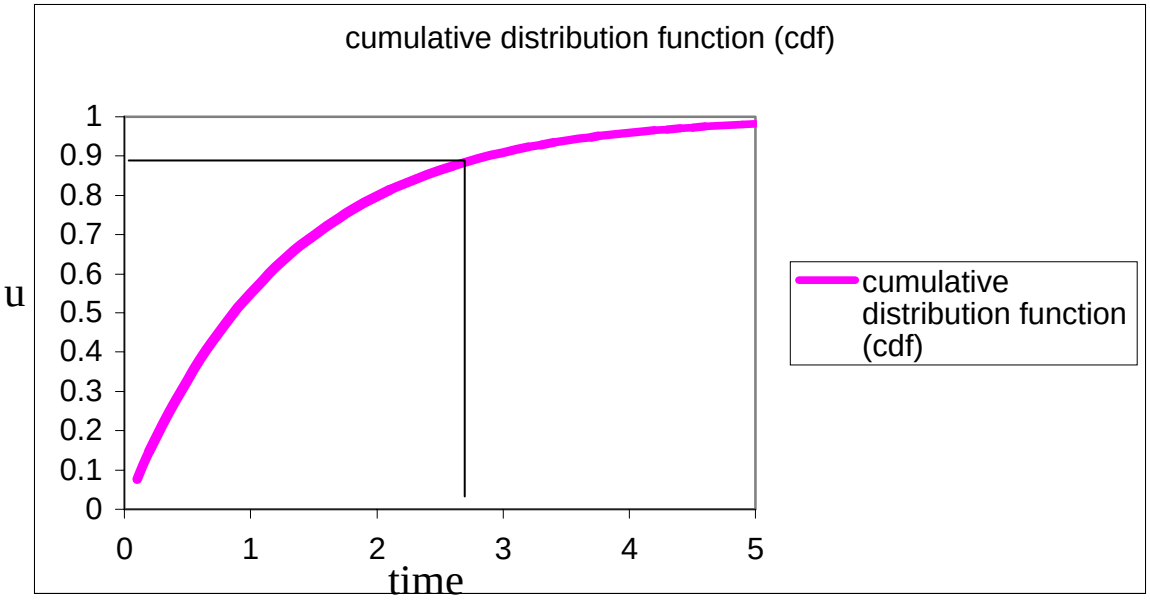
7. $f(x) = e^{g(x)}$

   $f'(x) = e^{g(x)}g'(x)$

8. $f(x) = g(x) + h(x)$

   $f'(x) = g'(x) + h'(x)$

We can generate uniform random variables using the "RAN#" function on our calculator, and then use the **cumulative distribution function** to transform them to their corresponding sojourn times (*t* values). We will see that the *t* values we generate in this way will be exponentially distributed, which means that a (standardized) histogram of them will approximate the exponential probability density function.



As you can see from the above schematic, around 90% of the *t* values generated will occur from 0 to a little less than 3 (2.878 to be more precise); and the remaining portion of *t*'s will be greater than this value. If we were to plot out these values, the histogram produced would approximate the probability density function in its shape, with most values small (i.e. less than 3) and few values (only about 10%) larger than 3.

If you are unsure of this, simply generate values for *u* using your calculator, calculate the corresponding *t* values, then create a histogram of those time values to prove to yourself that this indeed is the case (see below for the equation for determing *t* using the cumulative distribution function)

.

Simulating evolution of a nucleotide sequence using the cumulative distribution function

Next, we simulated the evolution of one nucleotide site over a period of time by first drawing random values for *u* (which range from 0 to 1), then deducing sojourn times (*t*) according to the cumulative distribution function.

First, we need to figure out how to obtain a sojourn time, *t*, given a uniform random number *u*. The first step is to equate *u* with F(t), which is the graphical equivalent of drawing a horizontal line from *u* on the y-axis to the right until it hits the F(t) curve.

$$u = 1 - e^{-\mu t}$$

Next, substract *u* from both sides to rearrange the equation and begin to isolate the *t* term

$$u-u = 1 - e^{-\mu t}-u$$

$$0 = 1 - e^{-\mu t} - u$$

Adding $e^{-\mu t}$ to both sides yields:

$$e^{-\mu t} = 1- u$$

Taking the natural log of both sides of the equation gets rid of the *e*:

$$\ln e^{-\mu t} = \ln (1-u)$$

$$-\mu t = \ln (1-u)$$

Dividing through by $-\mu$ completes the isolation of *t*:

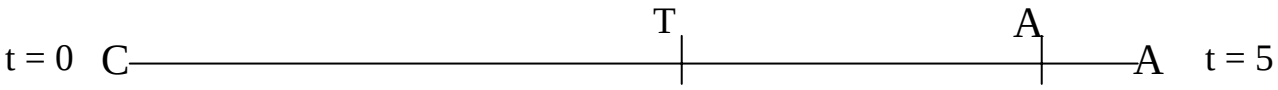
$$t = -\ln (1-u)/\mu$$

This formula represents the algebraic equivalent of transforming the value of **u** to the corresponding value of **t** as described by the CDF. So, once we have values of **u** we can calculate the corresponding values of **t,** given their relationship via the cumulative distribution function. We also need to specify a specific value for μ in order to actually generate any sojourn times, so let's just pick $\mu = 0.1$ arbitrarily for the purposes of this example.

Values for u	sojourn time t
0.269	3.13
0.131	1.40
0.449	5.96
0.711	12.41
0.781	15.19

We can use these values to simulate evolution of a single site in a nucleotide sequence. A sojourn time represents the time until the next disruption event. Thus after each sojourn time, we chose a base at random (each base has probability 0.25) to replace the current base.

Our final sequence looked like this, with a total of two disruption events from time 0 to time 5.



Note added by Paul: the amount of time represented by the branch (i.e. 5) was pulled out of the air just like the value of μ.

The portion of this branch from time 0 to the change marked by a “T” is the first sojourn time (3.13 time units). The second sojourn time (from T to A) has a total length of 1.4, but the total “waiting time” from the beginning of the branch until this second disruption (whereupon an A replaced the existing base T) is equal to 3.13 + 1.40, or 4.53.

Note by Paul: The final sojourn was 5.96 time units in length, and carried us beyond the end of the branch, which is why the base present at the end of the branch is an A.

We then used the remaining two sojourn times to simulate evolution at the second and third sites in our imaginary nucleotide sequence

Note by Paul: only one sojourn time was needed for each of next two sites because both are longer than 5. This means that no disruptions occurred along the branch, and hence the base at the end of the branch must have been equal to the base at the start.