3/2 PHYLOMATH LECTURE 6: Gamma Distributed Relative Rates

by Roberta Engel, 12 Mar 2004

Review

We started class with a review of the proportion of invariant sites (pinvar) style of rate heterogeneity. Sites either evolve at relative rate *r* or do not evolve at all (relative rate 0). The overall rate at which the sequences evolve is μ. We want a separate measure of the overall or average rate (μ) because then we can compare pinvar across different data sets.

site	1	2	3	4	5	…	n
relative rate	0μ	0μ	rμ	0μ	rμ		0μ

Some sites are freer to change than others (e.g. the 3rd site in an amino acid codon). We can incorporate rate heterogeneity into analyses or simulations to make a model more realistic.

We set E(*r*) = 1
= (0)(pinvar) + (r)(1-pinvar)

If we solve for *r* we can separate the pattern from the average rate of change (μ) and see that *r* is inversely related to pinvar.

$$r = 1/ (1\text{-pinvar})$$

When pinvar = 0 there is rate *homo*geneity. All the sites vary and they all evolve at the same relative rate (r = 1, rμ = μ for all sites).

When pinvar > 0 there is rate *hetero*geneity. As pinvar increases, fewer sites vary but those that do vary do so at a higher rate (r > 1 for sites having non-zero rate, rμ > μ).

Rate Heterogeneity as a Gamma Distribution

If we now think of relative rates as being Gamma(α, β) distributed, we can accommodate a range of relative rates. Two examples follow:

If α = 1, β= 1, the variance in relative rates equals 1.0

site	1	2	3	4	5	...
rel rate	.569	.194	.294	6.03	1.17	...

Some rates of change are slow (e.g. 0.194, 20%) and others are fast (e.g. 6.03, 600%). A variance of 1 means there is high rate heterogeneity. (see Figure 1)

If α = 200, β= 1/200, the variance is 1/200

site	1	2	3	4	5	...
rel rate	1.00	1.05	1.17	1.12	.86	....

r_1 r_2 r_3 r_4 r_5

In this case the relative rates are more similar and the range of values is not wide. A variance of 1/200 means there is low rate heterogeneity (or relative rate homogeneity). (see Figure 2)

In both examples, the mean of the relative rate values is 1 by design (again, the idea is to keep the average rate at which a gene is evolving separate from our measure of rate heterogeneity so that we can compare these measures across genes).

Determining Expected Value of Gamma Distributed Random Variable

In order to know how to set the shape (α) and scale (β) parameters of the Gamma distribution such that the mean is 1, we must first figure out the expected value of a Gamma(α, β) distribution. The probability density function (PDF) for the variable *R* is written as:

$$f(r) = \frac{r^{\alpha-1}e^{-r/\beta}}{\beta^{\alpha}\Gamma(\alpha)}$$

The density function integrates to 1.0 over $0 \rightarrow \infty$:

$$\int_0^{\infty} f(r)dr = 1.0$$

The numerator with the variable *r* is the real 'meat' of the density function. We may treat the denominator $\beta^{\alpha}\Gamma(\alpha)$ as a constant c since there is no *r* in the denominator:

$$\int_0^{\infty} \frac{r^{\alpha-1}e^{-r/\beta}}{c} dr = 1.0$$

Note that the constant is simply the scaling factor needed to make the density function integrate to 1.0. Thus, the following is true:

$$\int_0^{\infty} r^{\alpha-1}e^{-r/\beta}dr = c$$

This constant in the denominator consists of the product of β^{α} and the "gamma function" (not to be confused with the Gamma density function, of which it is a component!). The gamma function is symbolized $\Gamma(\alpha)$ and is described further below.

To say that the random variable R is Gamma distributed with shape parameter α and scale parameter β, write:

$R \sim \text{Gamma}(\alpha, \beta)$

α and β are parameters. α determines the shape of the density and β scales values up or down.
We can obtain the probability density of a particular value *r* of the random variable *R* by plugging-in values for α and β in the formula for $f(r)$. A particular *r* will be more or less probable depending on the values plugged-in. (This is like when we plug-in values for μ to get sojourn times.)

The gamma distribution has two basic shapes, an L-shaped curve (Figure 1) and a spike centered around a relative rate of 1 (Figure 2). When $\alpha \leq 1$, most of the relative rates are < 1 and only a few are >1. When $\alpha > 1$, there is not a broad range of rates rather the relative rates are close to 1. (Relative rate homogeneity is approached as α nears ∞.)

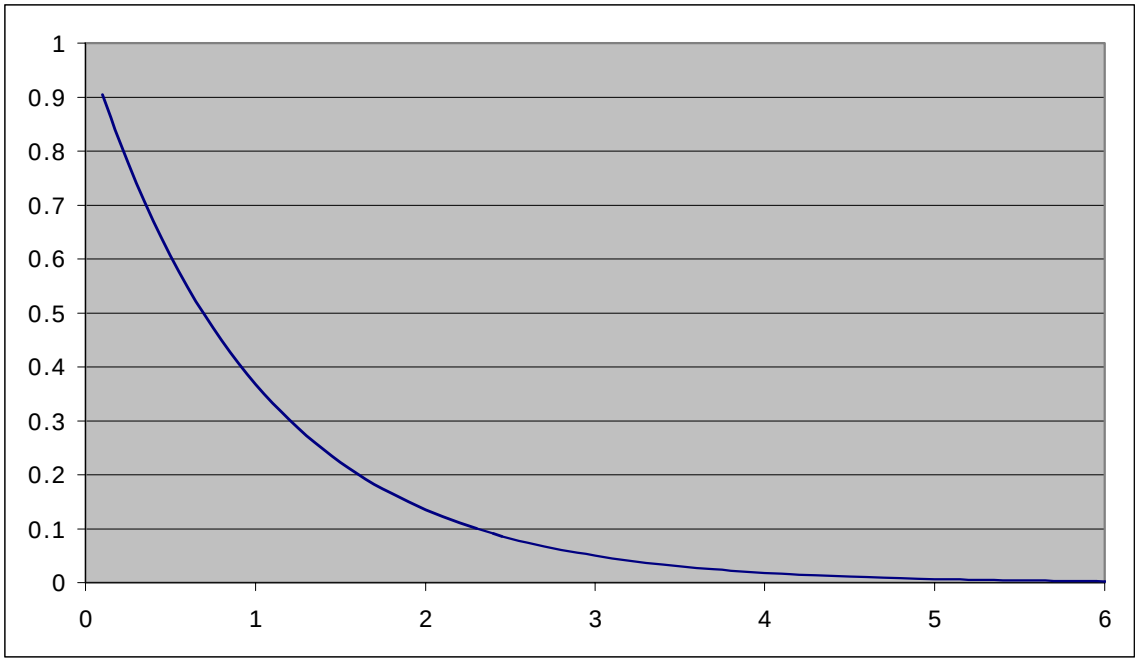

Figure 1 Basic Shape of Gamma Distribution when $\alpha = 1$. At r=0, the density is 1. When $\alpha < 1$, density becomes strongly L-shaped and the density is infinite at r=0.

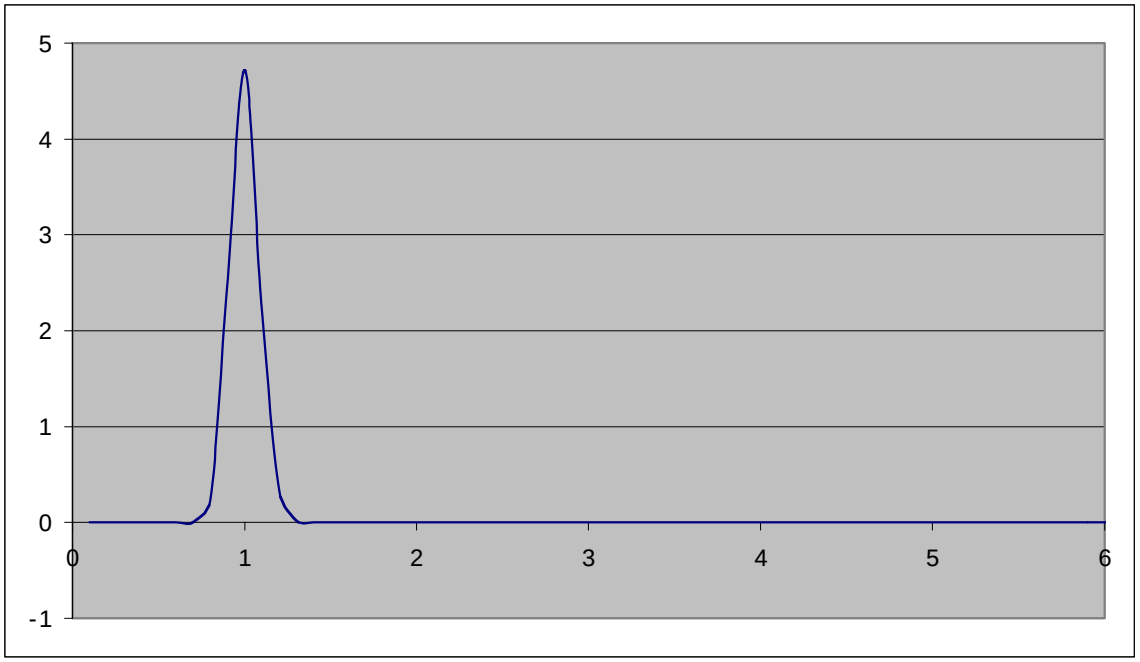

Figure 2 Basic Shape of Gamma Distribution when $\alpha > 1$ ($\alpha = 140$ in this case)

Recall last time with pinvar, we set:

$$E(r) = 1$$

Now, we will find E(r) when r is gamma distributed, then set this quantity equal to 1 in order to determine which combinations of α and β are suitable for modeling relative rates:

It is helpful to remember the equation:

$$\int_0^\infty f(r)dr = 1.0$$

The value of *r* can fall anywhere between 0 - ∞. It follows that:

$$E(r) = \int_0^\infty rf(r)dr$$

$$= \int_0^\infty r\left[\frac{r^{\alpha-1}e^{-r/\beta}}{\beta^\alpha \Gamma(\alpha)}\right]dr$$

$$= \frac{1}{\beta^{\alpha}\Gamma(\alpha)}\int_0^{\infty} r^{\alpha} e^{-r/\beta} dr$$

We factor the constant ($\beta^{\alpha}\Gamma(\alpha)$) outside the integral and try to make the integrand look like a gamma distribution. (Factoring out the constant from the integral is like factoring out a constant from a sum.)

Now we use Trick #1: let $\alpha = a - 1$ (or $a = \alpha + 1$) and multiply by 1 or $\frac{\beta^{a}\Gamma(a)}{\beta^{a}\Gamma(a)}$

$$\begin{aligned} E(R) &= \frac{1}{\beta^{\alpha}\Gamma(\alpha)}\int_0^{\infty} r^{\alpha} e^{-r/\beta} dr \\ &= \frac{1}{\beta^{\alpha}\Gamma(\alpha)}\int_0^{\infty} r^{a-1} e^{-r/\beta} dr \\ &= \frac{1}{\beta^{\alpha}\Gamma(\alpha)}\int_0^{\infty} \left(\frac{\beta^{a}\Gamma(a)}{\beta^{a}\Gamma(a)}\right)\left(r^{a-1} e^{-r/\beta}\right) dr \\ &= \frac{\beta^{a}\Gamma(a)}{\beta^{\alpha}\Gamma(\alpha)}\int_0^{\infty} \frac{r^{a-1} e^{-r/\beta}}{\beta^{a}\Gamma(a)} dr \end{aligned}$$

None of these manipulations actually changes anything (r^{a-1} equals r^{α}, and multiplying by 1 is always acceptable), but these operations set things up so that we can rid ourselves of that ugly integral:

Noting that $\int_0^{\infty} \frac{r^{a-1} e^{-r/\beta}}{\beta^{a}\Gamma(a)} dr = 1.0$, we have, simply, $E\{R\} = \frac{\beta^{a}\Gamma(a)}{\beta^{\alpha}\Gamma(\alpha)}$

Next we will use Trick#2: $\Gamma(x) = (x-1)\,\Gamma(x-1)$ and thus $\Gamma(x+1) = x\Gamma(x)$ to simplify:

$$= \frac{\beta^{a}\Gamma(a)}{\beta^{\alpha}\Gamma(\alpha)}$$

$$= \frac{\beta^{\alpha+1}\Gamma(\alpha+1)}{\beta^{\alpha}\Gamma(\alpha)}$$

$$= \frac{\beta^{\alpha+1}\alpha\Gamma(\alpha)}{\beta^{\alpha}\Gamma(\alpha)}$$

$$= \frac{\beta^{\alpha+1}\alpha}{\beta^{\alpha}}$$

$$= \alpha\beta$$

So, setting the expected value (i.e. $\alpha\beta$) to 1, we can express β in terms of α:

If $\alpha\beta = 1$, then $\beta = \frac{1}{\alpha}$

We never hear about the scale parameter of the Gamma distribution in phylogenetic analyses because the scale parameter needs to be $1/\alpha$ in order for the mean to be 1. Thus, the shape parameter is everything.

The variance of the gamma density function is

$\text{Var}(R) = \alpha\beta^2 = \frac{\alpha}{\alpha^2} = \frac{1}{\alpha}$

We did not work this out in class, but the same techniques are used to figure this out. As α increases (e.g. 200), the variance is small (e.g. 1/200) and, conversely, as α decreases (e.g. 1/100), variance (i.e. rate heterogeneity) is large (100).

A Brief Aside to Make Connection between Gamma Distribution and Sojourn Time Distribution

A natural use of gamma distribution equations is to describe waiting times. They are similar concepts and share the same thought processes. In fact, the exponential distribution is a special case (when $\alpha = 1$) in the gamma distribution family. The sojourn time is defined to be the waiting time to the first "event" (call it w_1). The "event" we have been tracking have been either "disruptions" (which occur at rate μ) or substitutions (which occur at rate α, but note that this α has nothing to do with the gamma shape parameter! The same symbols are unfortunately used all over the place). The waiting time to the second disruption (w_2) has a Gamma distribution with shape parameter equal to 2). The third waiting time (w_3) is gamma distributed with shape parameter equal to 3. In general, Gamma (α, β) = distribution of α^{th} waiting time when mean sojourn time = β. Exponential distribution = Gamma(1,β) = distribution of 1^{st} waiting time when mean sojourn time = β.

Here we recall that sojourn time (t) is distributed as the exponential distribution with parameter μ:

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

This is the gamma distribution in disguise.

$$= \frac{t^{\alpha-1}e^{-t/\beta}}{\beta^{\alpha}\Gamma(\alpha)} \qquad \alpha = 1, \beta = \frac{1}{\mu}$$

Because $\Gamma(1) = 1$ then

$$= \frac{t^{0}e^{-t\mu}}{\frac{1}{\mu}\,\Gamma(1)}$$

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

$$E(t) = \alpha\,\beta = \frac{1}{\mu}$$

Thus, the expected sojourn time is $1/\mu$ when disruptions occur at rate μ. This makes sense: if disruptions rarely occur, sojourn times are long. If disruptions are occurring at a high rate, sojourn times will be shorter.

Discrete Gamma Distribution

We do not use the full gamma distribution to get continuous relative rates in phylogenetic analyses because the computation time is too intensive. Instead, we use a practical model, known as discrete gamma distribution (Yang 1994).

The gamma distribution changes as we vary relative rate heterogeneity. Note the distributions on the three graphs below:

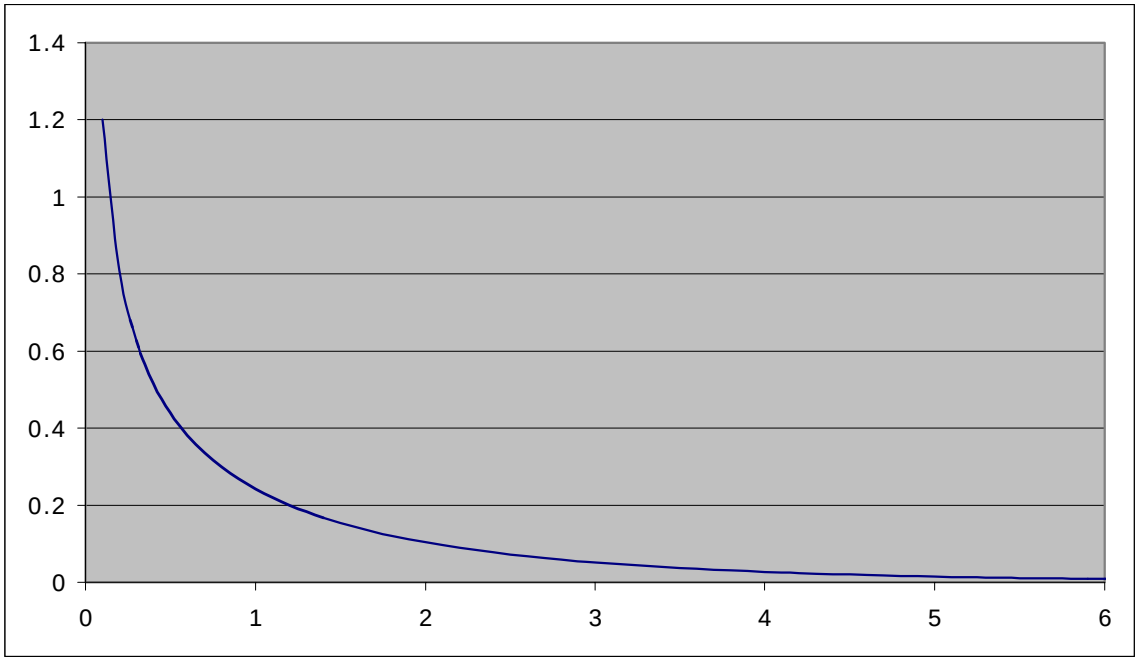

Figure 3 Gamma Distribution $\alpha = 0.5$

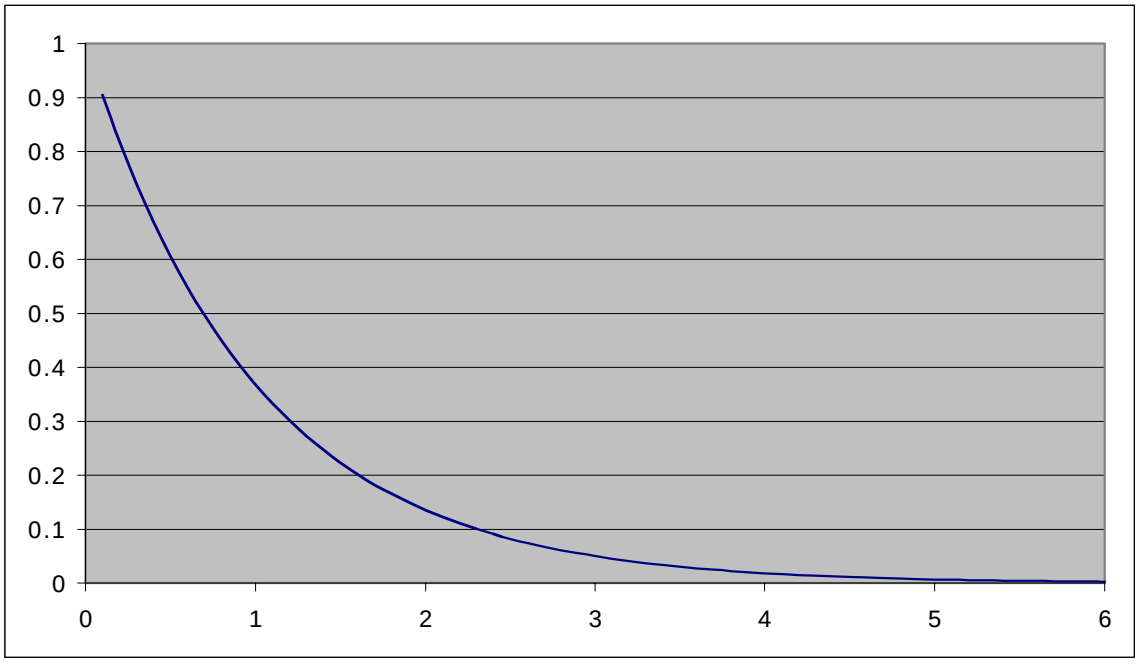

Figure 4 Gamma Distribution $\alpha = 1$

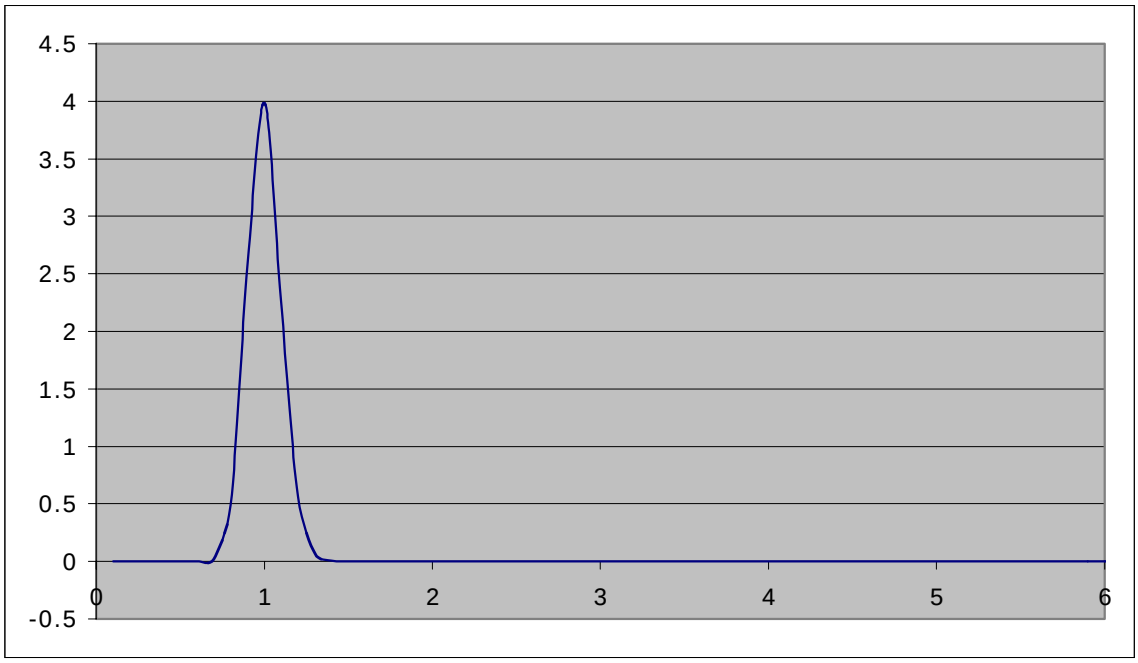

Figure 5 Gamma Distribution $\alpha = 100$

We divided the distribution into a number of categories (ncat). Each chunk has the same area (1/ ncat). Note the lines are not evenly spaced on x-axis. The mean of the relative rate for each category *i* is the representative relative rate, r_i.

site	1	2	3	4	5	n
rel rates	r_1	r_1	r_3	r_4	r_2	r_1

The result is not a very fine-grained model of rate heterogeneity because we use a representative relative rate from each rate class (i.e. we are reducing our choices for relative rates from infinity down to just ncat). If we use more categories, it leads to smaller chunks, and the relative rates are closer to the values they would be if we used a continuous gamma distribution. As we increase the number of categories, we increase the amount of time needed to evaluate likelihoods of trees. The default value tends to be 4, and up to 8 categories works well for data sets, but beyond this the results do not usually pay off due to the time cost.

Typically, if we decide to employ discrete gamma rate heterogeneity, we rely on a program such as PAUP* to estimate the shape parameter for us from our data. Given a particular value for the shape parameter, you can type "gammaplot shape = (value for α)" into PAUP* to generate a crude plot of the distribution. The default in PAUP* is α =0.5. PAUP* also provides the cutoff points for each rate category, as well as the representative (mean) rate for each category. It is important for us to know how PAUP* calculates the values, and while calculating the values completely by hand is out of the question, Excel combined with some knowledge of Gamma distributions can be used to see where these numbers come from.

We use the CDF (cumulative density function, F(r)) to calculate the integral of the density function up to a specified point.

$$F(a) = \int_0^a f(r)dr$$

F(a) = 1/4
F(b) = 1/2
F(c) = 3/4

Let's assume the shape parameter is 0.5 and the scale parameter is 2.0 for the following examples. We can plug-in a relative rate value (e.g. r=1.12, and the Excel function GAMMADIST(1.12, 0.5, 2.0, true) can be used to evaluate the integral represented by F(r) (which is the area under the density curve from 0 up to 1.12). The last argument "true" tells Excel that we want the CDF rather than the PDF of the Gamma distribution. Note that the shape parameter is the second argument, and the scale parameter is the third argument to this function.

Conversely, we can plug F(r) area values (e.g. F(r) = 0.25) in, and the Excel GAMMAINV function will calculate the proper value of r such that the area under the

curve from 0 up to r equals 0.25. For example, GAMMAINV(0.25, 0.5, 2.0) returns the value of the first cutoff point (a), GAMMAINV(0.5, 0.5, 2.0) returns the second cutoff point (b), and GAMMAINV(0.75, 0.5, 2.0) returns the third cutoff point (c). The fourth cutoff point is of course infinity.

How to get the Mean for one rate category

Let's use the discrete distribution in Figure 6 to illustrate what we are trying to do. The vertical axis gives the frequency of the numbers on the horizontal axis. The total number of observations is 24.

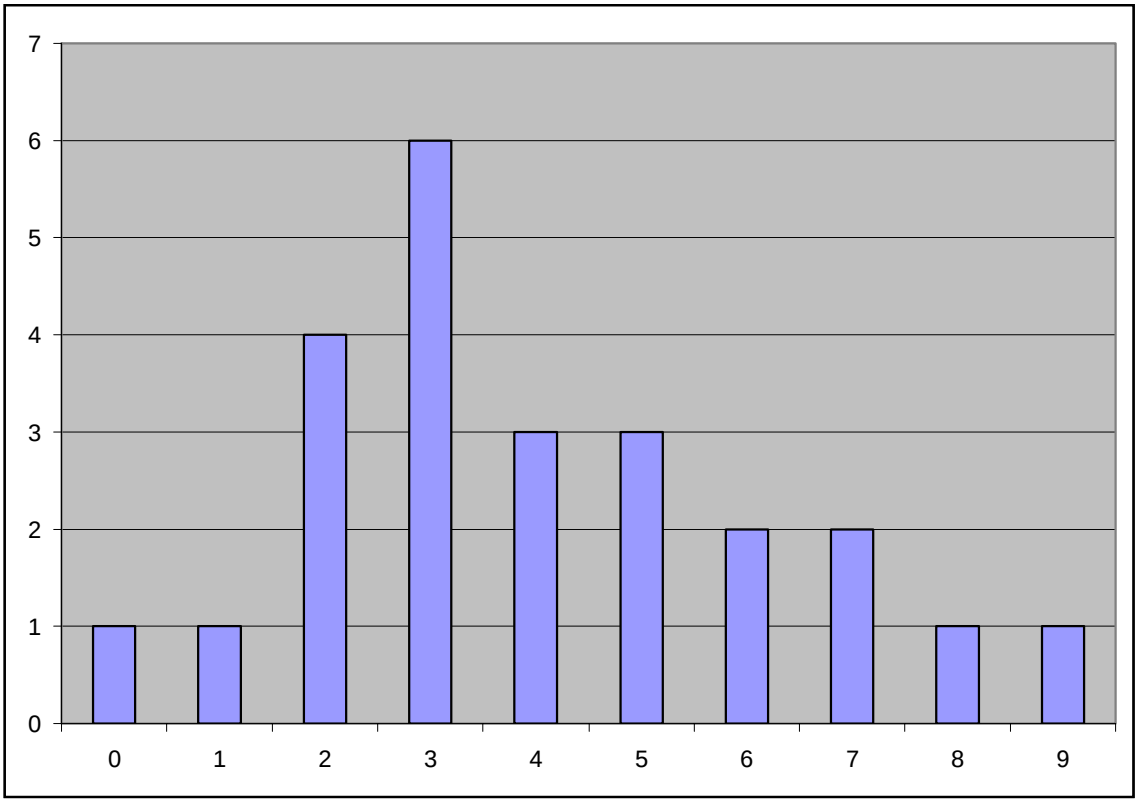

Figure 6 Discrete Distribution

The overall mean may be calculated as follows:

$$0\left(\frac{1}{24}\right)+1\left(\frac{1}{24}\right)+2\left(\frac{2}{24}\right)+3\left(\frac{6}{24}\right)+4\left(\frac{3}{24}\right)+5\left(\frac{3}{24}\right)+6\left(\frac{2}{24}\right)+7\left(\frac{2}{24}\right)+8\left(\frac{1}{24}\right)+9\left(\frac{1}{24}\right)$$

or we can write this as just

$$= \frac{(0)(1)+(1)(1)+(2)(4)+\ldots+(9)(1)}{24}$$

(The observed probability of each value is frequency/24, and the mean is the expected value computed with these observed probabilities)

We can also calculate the mean of the chunks, considering one chunk at a time.
So the mean of chunk 1 is:

$$= (0)\frac{1}{6} + (1)\frac{1}{6} + (2)\frac{4}{6}$$

$$= \frac{1}{6} + \frac{8}{6} = \frac{9}{6} = \frac{3}{2}$$

or:

$$\frac{(0)(1) + (1)(1) + (2)(4)}{6} = \frac{1+8}{6} = \frac{9}{6} = \frac{3}{2}$$

Note that the numerator is the sum of values times frequencies, and the denominator is the sum of only the frequencies. Reverting back to the problem at hand, where sums become integrals and probabilities become $f(r)\,dr$, the mean of the 1st chunk may be written in terms of the PDF as:

$$= \frac{\int_0^a rf(r)dr}{\int_0^a f(r)dr}$$

$$= \frac{\int_0^a rf(r)dr}{\frac{1}{4}} \qquad \text{(because F(a) = 1/4)}$$

$$= \int_0^a r\left[\frac{r^{\alpha-1}e^{-r/\beta}}{\beta^{\alpha}\Gamma(\alpha)}\right]dr$$

We can get parts of this equation to look like a gamma distribution by using our previous bag of tricks:

$$= \int_0^a \frac{r^{\alpha}e^{-r/\beta}}{\beta^{\alpha}\Gamma(\alpha)}dr$$

$$= \frac{\beta^{x}\Gamma(x)}{\beta^{x}\Gamma(x)}\int_0^a \left[\frac{r^{x-1}e^{-r/\beta}}{\beta^{x}\Gamma(x)}\right]dr$$

Note that we cannot simply forget about that integral this time because the upper limit is a (not infinity) so it doesn't equal 1. We can use Excel to compute this integral, however. This integral is the CDF of a Gamma distribution in which the shape parameter is *x* rather than α, where $x = \alpha + 1$.

Use Excel gammadist (a, α + 1, β, true) to evaluate this integral, where a = boundary; α + 1 = shape; β = scale; true means CDF)

Noting that $\frac{\beta^x \Gamma(x)}{\beta^\alpha \Gamma(\alpha)} = \alpha\beta$ as before, we get the mean of the first category in Excel by:

$4\alpha\beta$ gammadist (a, α + 1, β, true)

The mean values for category 1 are comparable:
PAUP* = 0.03338775
Excel = 0.03338767

To find the mean for the 2nd category, we want the integral to go from a (first cutoff point) to b (second cutoff point):

$$r_2 = \frac{\int_a^b r \frac{r^{\alpha-1} e^{-r/\beta}}{\beta^\alpha \Gamma(\alpha)} dr}{\frac{1}{4}} = 4\int_a^b \frac{r^\alpha e^{-r/\beta}}{\beta^\alpha \Gamma(\alpha)} dr = 4\alpha\beta \int_a^b \frac{r^{x-1} e^{-r/\beta}}{\beta^x \Gamma(x)} dr$$

In Excel, find the integral from 0 to b then subtract the integral from 0 to a:

$r_2 = 4\alpha\beta$ [gammadist (b, α + 1, β, true) - gammadist (a, α + 1, β, true)]

For the 3rd category:

$r_3 = 4\alpha\beta$ [gammadist (c, α + 1, β, true) - gammadist (b, α + 1, β, true)]

And for the final (4th) category:

$r_3 = 4\alpha\beta$ [1 - gammadist (c, α + 1, β, true)]

Gamma Function is a Generalized Factorial

Gamma function is a generalized factorial so:

$\Gamma(n + 1) = n!$

To see why this is true, apply the definition of the gamma function repeatedly, plugging in $\Gamma(1) = 1$ at the end:

$$\begin{aligned}\Gamma(n+1) &= n\Gamma(n)\\ &= n(n-1)\Gamma(n-1)\\ &= n(n-1)(n-2)\Gamma(n-2)\\ &\\ &= n(n-1)(n-2)\ldots(2)(1)\Gamma(1)\\ &= n(n-1)(n-2)...(2)(1)(1)\\ &= n!\end{aligned}$$

It is a *generalized* factorial because you can plug fractional numbers into the gamma function, whereas factorials only are defined for integer values.

The number n! is too large for most computers to handle when n is very large (try 600! on your calculator for example), so to get around this you can calculate $\ln\Gamma(n + 1)$, which equals ln(n!). For most situations, knowing the natural log of the factorial is all that is necessary.

Note added by Paul (this example was not given in lecture, but seems appropriate here): For example, consider computing the binomial coefficient $\binom{n}{y}$ where n=600 and y=100.

Trying to do this directly

$$\binom{n}{y} = \frac{n!}{y!(n-y)!} = \frac{600!}{100!\,500!}$$

is not possible (100! is already 9.3×10^{157}, and 600! is just too big a number for Excel to handle). Computing the natural log of $\binom{n}{y}$ is fairly easy using the GAMMALN function in Excel:

$$\ln\left(\frac{\Gamma(601)}{\Gamma(101)\,\Gamma(501)}\right) = \ln\Gamma(601) - \ln\Gamma(101) - \ln\Gamma(501) = 267.2055014$$

The desired binomial coefficient can now be calculatedas $e^{267.2055014} = 1.11\times10^{116}$.

References

Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. Journal of Molecular Evolution 39:306-314.

Yang, Z. 1993. Maximumlikelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Molecular Biology and Evolution 10:1396-1401.