Pinvar approach

- Unlike the site-specific rates approach, this approach does not require you to assign sites to rate categories
- Assumes there are only two classes of sites:
  - invariable sites (evolve at relative rate 0)
  - variable sites (evolves at relative rate *r*)
- Remarks:
  - mean of relative rates = $(p_{invar})(0) + (1-p_{invar})(r) = 1$
  - this means that $r = 1/(1-p_{invar})$
  - if all sites are variable, $p_{invar} = 0$ and $r = 1$

- **Constant site** – a site in which all of the taxa display the same character state.
- **Invariable site** – a site in which only one character state is allowed. A site that cannot change state.

All invariable sites are constant, but not all constant sites have to be invariable.

$$
\begin{aligned}
\Pr(i \to i|\ \text{invariable}) &= \frac{1}{4} + \frac{3}{4}\, e^{\frac{-40\nu}{3}} \\
&= \frac{1}{4} + \frac{3}{4}\, e^{0} \\
&= 1 \\
\Pr(i \to j|\ \text{invariable}) &= \frac{1}{4} - \frac{1}{4}\, e^{\frac{-40\nu}{3}} \\
&= 0
\end{aligned}
$$

A site's likelihood under the JC+ I model

x_i is the data pattern for site i. General form:

$$\Pr(x_i|\text{JC+I}) = p_{\text{inv}} \Pr(x_i|\text{ inv}) + (1 - p_{\text{inv}}) \Pr\left(x_i|\text{JC}, \frac{\boldsymbol{\nu}}{1 - p_{\text{inv}}}\right)$$

If x_i is a variable site:

$$\Pr(x_i|\text{JC+I}) = (1 - p_{\text{inv}}) \Pr\left(x_i|\text{JC}, \frac{\boldsymbol{\nu}}{1 - p_{\text{inv}}}\right)$$

If x_i is a constant site:

$$\Pr(x_i|\text{JC+I}) = p_{\text{inv}} \Pr(x_i|\text{ inv}) + (1 - p_{\text{inv}}) \Pr\left(x_i|\text{JC}, \frac{\boldsymbol{\nu}}{1 - p_{\text{inv}}}\right)$$

Why $\frac{\nu}{1-p_{\mathsf{inv}}}$?

We want the mean rate of change to be 1.0 over all sites (so we can interpret the branch lengths in terms of the expected # of changes per site).

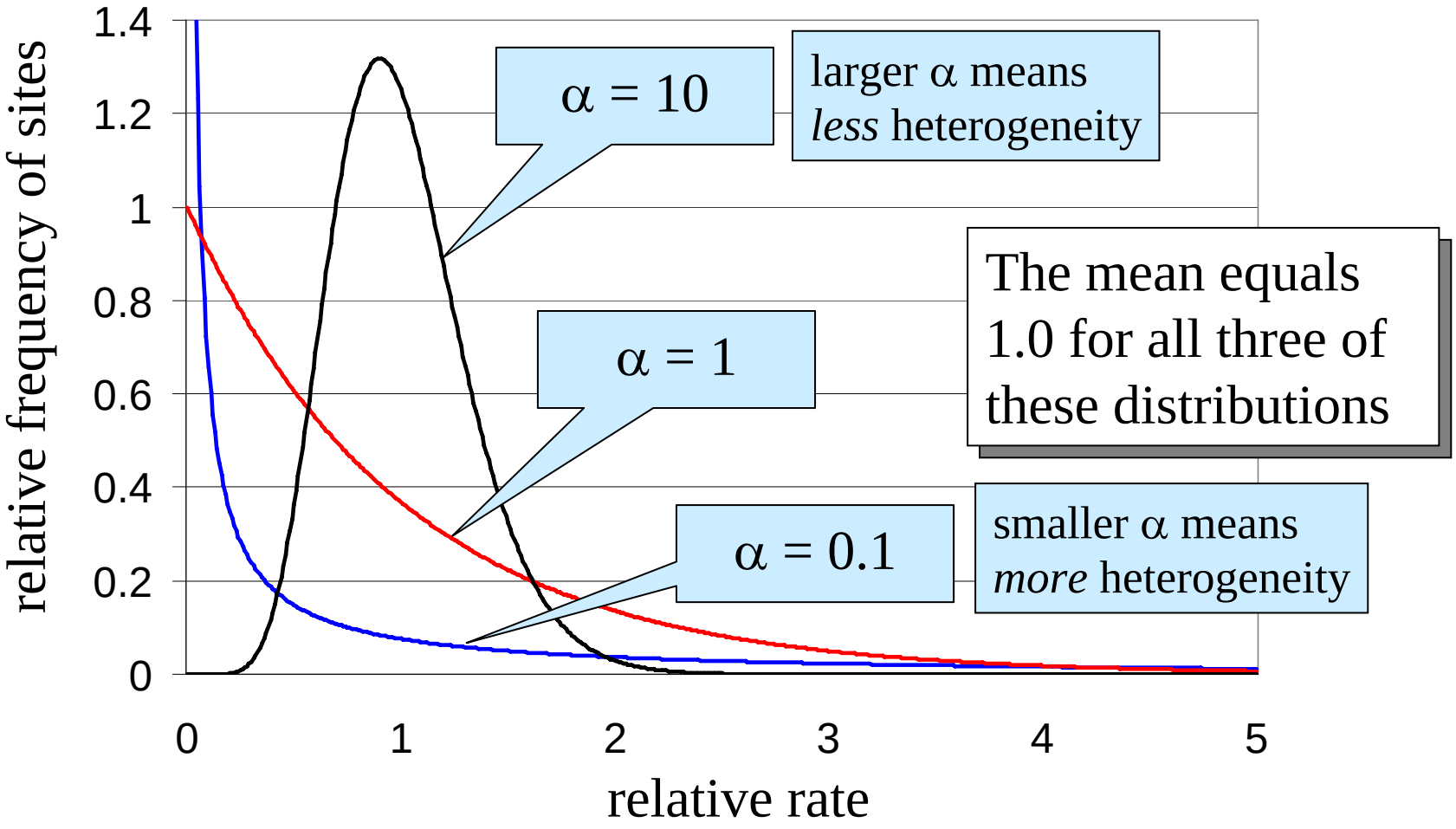
If r is the rate of change for the variable sites then:

$$\begin{aligned} 1 &= 0p_{\mathsf{inv}} + r\left(1 - p_{\mathsf{inv}}\right) \\ &= r\left(1 - p_{\mathsf{inv}}\right) \\ r &= \frac{1}{1 - p_{\mathsf{inv}}} \end{aligned}$$

Variable (but unknown) rates

- We expect more "shades of grey" rather than the on-or-off view of the pInvar model.
- *a priori* we do not know which sites are fast and which are slow
- We may be able to characterize the *distribution* of rates across sites – high variance or low variance.

Gamma distributions



Gamma distribution

$$
\begin{aligned}
f(r) &= \frac{r^{\alpha-1}\beta^{\alpha}e^{-\beta r}}{\Gamma(\alpha)} \\
\text{mean} &= \alpha/\beta \\
\text{mean (in phylogenetics)} &= 1 \\
\text{(in phylogenetics) } \beta &= \alpha \\
\text{variance} &= \alpha/\beta^2 \\
\text{variance (in phylogenetics)} &= 1/\alpha
\end{aligned}
$$

Using Gamma-distributed rates across sites

- We usually use a discretized version of the gamma with 4-8 categories (the computation time increases linearly with the number of categories).

$$\Pr(x_i|JC+G) = \sum_{j}^{\text{ncat}} \Pr(x_i|JC, r_j\boldsymbol{\nu}) \Pr(r_j)$$

where:

$$\sum_{j}^{\text{ncat}} r_j \Pr(r_j) = 1$$

Discrete gamma (continued)

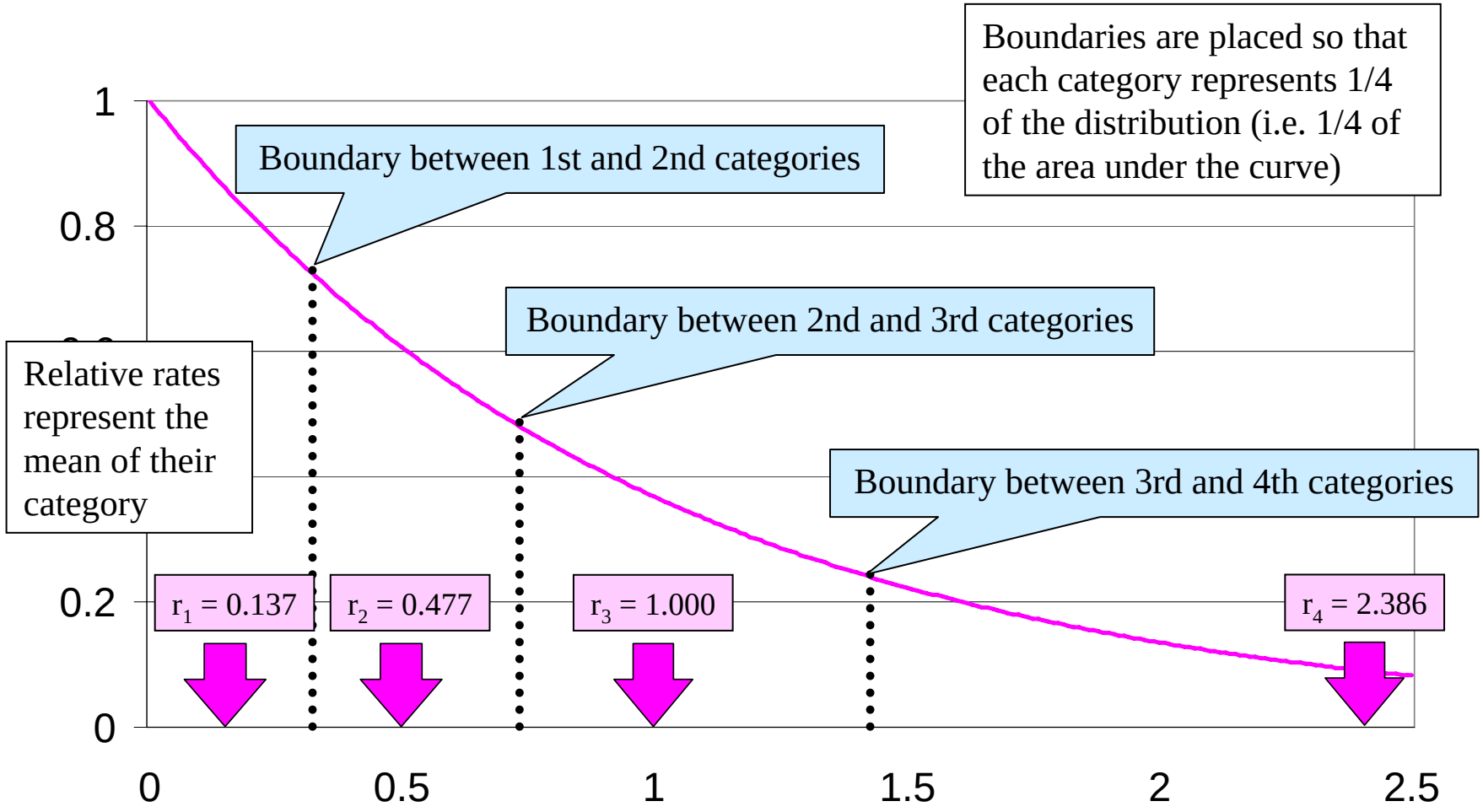
We "break up" the continuous gamma into intervals each of which has an equal probability, and use the mean rate within each interval as the representative rate for that rate category:

$$\Pr(r_j) = \frac{1}{\text{ncat}}$$

So:

$$\Pr(x_i|JC + G) = \frac{1}{\text{ncat}} \sum_{j}^{\text{ncat}} \Pr(x_i|JC, r_j\boldsymbol{\nu})$$

Relative rates in 4-category case



Discrete gamma rate heterogeneity in PAUP*

To use gamma distributed rates with 4 categories:

```
lset rates=gamma ncat=4;
```

To estimate the shape parameter:

```
lset shape=estimate;
```

To combine pinvar with gamma:

```
lset rates=gamma shape=0.2 pinvar=0.4;
```

Note: `estimate`, `previous`, or a specific value can be specified for both `shape` and `pinvar`

Rate homogeneity in PAUP*

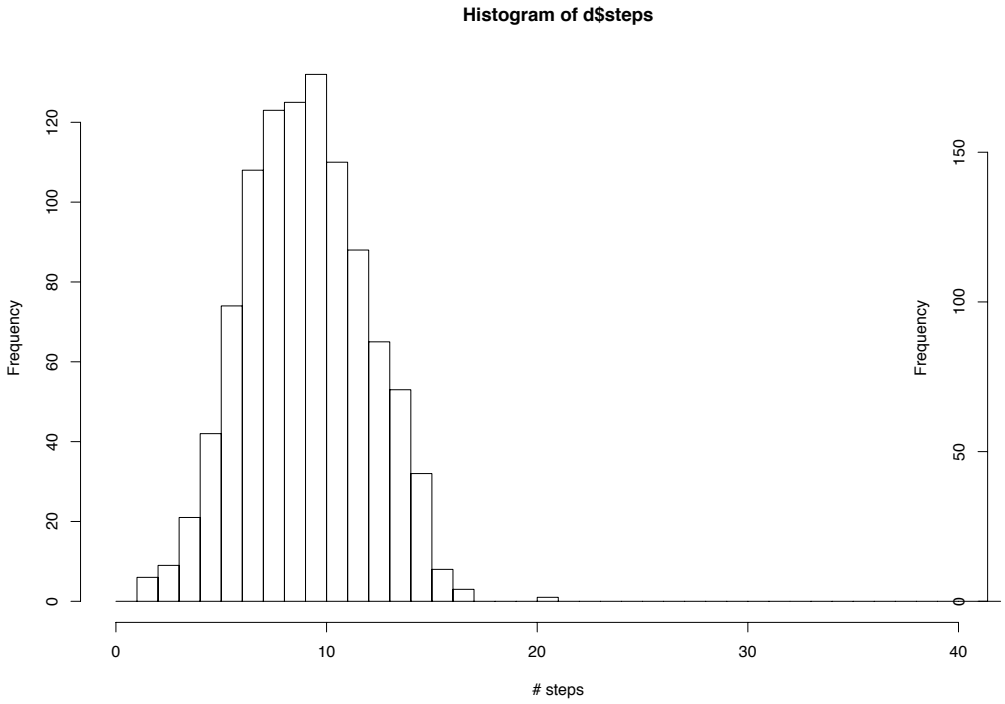
Just tell PAUP* that you want all rates to be equal and that you want all sites to be allowed to vary:

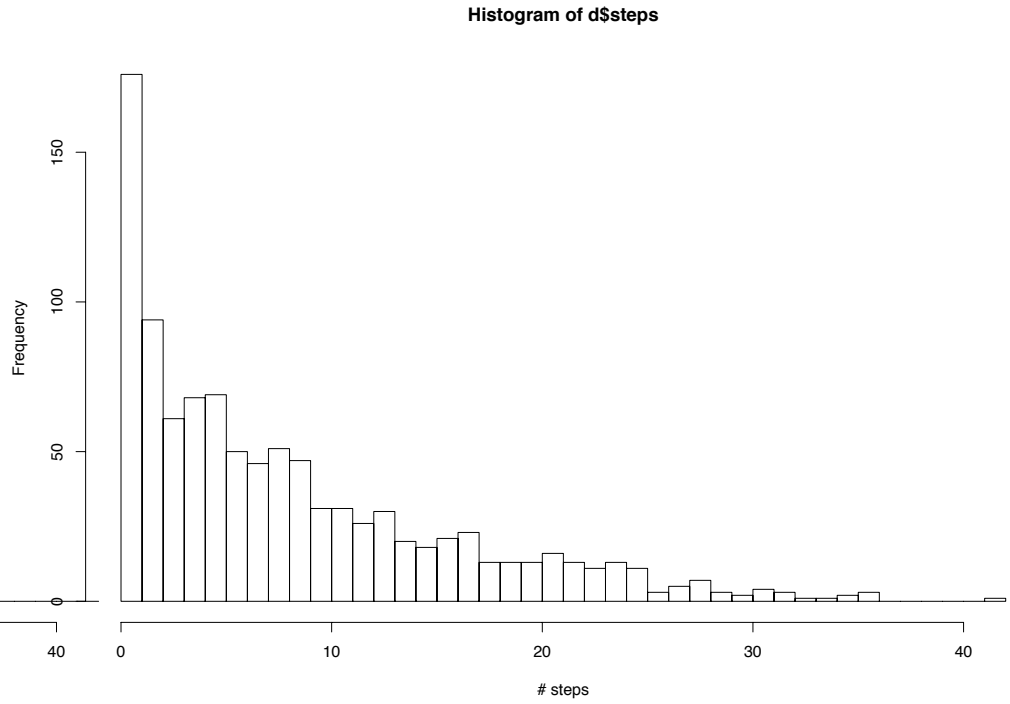
```
lset rates=equal pinvar=0;
```

Note: these are the default settings, but it is useful to know how to go back to rate homogeneity after you have experimented with rate heterogeneity!

Rate heterogeneity summary

1. among-character rate heterogeneity is pervasive, and detectable;
2. failure to account for it can lead to biased branch length estimates (and hence incorrect tree inference);
3. distance corrections can use estimates of rate heterogeneity (but this introduces a lot of variance);
4. recognizing fast characters can be thought of as "downweighting" them;


Histogram of d$steps
Frequency
0
20
40
60
80
100
120
0
10
20
30
40
steps


Histogram of d$steps
Frequency
0
50
100
150
0
10
20
30
40
steps

Rates of evolution and the "reliability" of characters

Character fit on a 130 taxon tree (simulated with rate heterogeneity):

	Char. 67		Char. 882	
	Tree 1	Tree 2	Tree 1	Tree 2
Pscore	4	5	31	30

Based on these 2 characters, both trees have 35 steps.

Rates of evolution and the "reliability" of characters (continued)

Character fit on a 130 taxon tree (simulated with rate heterogeneity):

	Char. 67		Char. 882	
	Tree 1	Tree 2	Tree 1	Tree 2
Pscore	4	5	31	30
K2P lnL	-23.92	-26.13	-117.22	-116.72

Preference for tree 1 (based on these two characters alone).

$\Delta \ln L = 1.7$ under K2P.

Rates of evolution and the "reliability" of characters (continued)

Character fit on a 130 taxon tree (simulated with rate heterogeneity):

		Char. 67		Char. 882	
		Tree 1	Tree 2	Tree 1	Tree 2
Pscore		4	5	31	30
K2P lnL		-23.92	-26.13	-117.22	-116.72
K2P+G lnL		-23.538	-26.348	-98.46	-98.1
categ. %	1	14.2%	3.94%	0%	0%
	2	80.4%	84.2%	0%	0%
	3	5.3%	11.84%	0.000017%	0.000019%
	4	0.000039%	0.0002 %	99.999983%	99.999981%

Stronger preference for tree 1 (based on these two characters alone).

$\Delta \ln L = 2.45$ under K2P+ Gamma.

Rate heterogeneity and parsimony

1. Successive weighting of Farris (1989): iterative reweighting by the rescaled consistency index on the best tree:

   Consistency index: $CI = \frac{\text{min. \# steps}}{\text{obs. \# steps}}$

   Retention index: $RI = \frac{\text{max. \# steps}-\text{obs. \# steps}}{\text{max. \# steps}-\text{min. \# steps}}$

   Rescaled consistency index: $RC = (CI)(RI)$

2. Implied weights of Goloboff (1993):

   $$\frac{K}{K+\text{obs. \# steps}-\text{min. \# steps}}$$

References

Goloboff, P. (1993). Estimating character weights during tree search. *Cladistics*, 9(1):83–91.