# On the size distribution of live genera.

William J. Reed
Department of Mathematics and Statistics,
University of Victoria,
Victoria, British Columbia,
CANADA V8W 3P4
reed@math.uvic.ca
and
Barry D. Hughes
Department of Mathematics and Statistics,
University of Melbourne,
Parkville, Victoria 3010,
AUSTRALIA
hughes@ms.unimelb.edu.au

June, 2001.

**Abstract**

This article deals with the theoretical size (number of species) distribution of live genera, arising from a simple model of macroevolution in which speciations and extinctions are assumed to occur independently and at random, and in which new genera are formed by the random splitting of existing genera. Mathematically the distribution is that of the state of a homogeneous birth-and-death process after an exponentially distributed time. An ordinary differential equation for the generating function of the distribution is derived and solved and a recurrence relation for computing the probabilities in the distribution presented. Some properties of the distribution, including asymptotic behaviour, are examined and the distribution of the time since establishment of a genus of a given size derived. Fitting the distribution to empirical taxon size distributions by maximum likelihood is discussed and two examples are presented

**Keywords:** Birth-and-death process, taxon size distribution, asymptotic methods, Yule model.

# 1 Introduction.

Empirical abundance distributions of the numbers of species per genus (and at higher taxonomic levels of numbers of sub-taxa per taxon) reveal a considerable degree of regularity: they are all extremely long-tailed; they have their mode at one (*i.e.* the monospecific genus is the most frequent); and when the frequency of genera of size $n$ is plotted against genus size $n$ on logarithmic axes, the points fall close to a straight line, at least for genera up to a certain size, suggesting power-law behaviour in the distribution. These commonalities appear to hold both for living taxa and for fossil taxa and have been recognized for a long time (Yule, 1924; Corbet, 1942). Burlando (1990, 1993) gives many examples of abundance distributions exhibiting these properties, and speculates that the power-law behaviour results from fractal dynamics in the underlying evolutionary process.

Yule (1942) presented a mathematical theory of macroevolution to explain observed abundance distributions, introducing and developing the theory of the eponymous *Yule process* (or homogeneous pure birth process). He derived the probability distribution of the number of species in a genus at a fixed time after its establishment, and under the assumption that genera are formed in a process stochastically similar to that in which species are formed, derived the distribution of genus sizes over genera in a family. This is now called the *Yule distribution*, and is known to exhibit power-law behaviour asymptotically, and to follow close to a power-law over the whole of its range. Yule however observed that in the extreme upper tail (genus sizes greater than about thirty) empirical distributions seemed to decay faster than a power-law, and he claimed that this was due to finite time effects - *i.e.* as a result of the fact that no genus in a given family could have been in existence for longer than a finite time, determined by the time of origin of the family.

In his model Yule ignored extinctions, claiming that in the main these have been due cataclysmic events. Yule's model can thus be thought of as a model of adaptive radiation. An obvious way to extend the model is to allow for random individual extinctions in addition to speciations, employing a homogeneous birth and death process and thus leading to a model of neutral macroevolution. This has been proposed by Raup (1985), who in an appendix presented standard mathematical results on such processes. He did not however give results analogous to Yule's for the size of a living genus, possibly because his main interest is in paleobiology [1]. An alternative approach was adopted by Chu and Adami (1999), who used a Galton-Watson branching process model, from which they derived an iterative scheme for determining taxon abundance distributions. However the discrete nature of

[1] The size of a fossil taxon is the number of sub-taxa which have existed in the taxon at any time prior to extinction, which, of course, if individual non-catastrophic extinctions are possible, is not the same as the number existing at the final time of extinction. We consider the size distribution of fossil taxa in a forthcoming paper (Reed & Hughes, 2001)

time in their model, and the fact that at each time step a sub-taxon can give rise to zero, one, two, ... *etc.* new sub-taxa renders it less realistic than the birth-and-death process model proposed by Raup. Kemp (1995) considered a variant of the Yule model in which the task of the taxonomist is modelled as a birth and death process ('lumping' two species corresponding to a death and 'splitting' a species to a birth) and derived logarithmic and polylogarithmic distributions for genus size.

In this paper we consider the birth and death process model proposed by Raup. It can be thought of as a null model in which speciations and extinctions occur independently and at random. Thus it does not include the possibility of episodic mass extinctions caused by cataclysmic events (Raup, 1985, 1991). We derive results for the size distribution of living genera and for the distribution of time since establishment for genera of a given size. The properties of the distributions are examined, and the statistical issues of fitting the abundance distribution to empirical data considered.

## 2 A null model for speciations and extinctions.

Consider a genus which begins with one species (or more generally a taxon with one member) at time $t = 0$. Suppose that in time $(t, t+h)$ there is a probability $\lambda h + o(h)$ that any given species may split into two distinct species (a speciation), and a probability $\mu h + o(h)$ that the individual species alone becomes extinct. Suppose further that all speciations and individual extinctions are independent. Let $N_t$ denote the number of species currently alive at time $t$ and let

$$p_n(t) = \Pr\{N_t = n\}. \tag{1}$$

Then for $h > 0$,

$$\begin{aligned} p_n(t+h) &= [1 - n(\lambda+\mu)h + o(h)]p_n(t) \\ &\quad +[\lambda h + o(h)](n-1)p_{n-1}(t) \\ &\quad +[\mu h + o(h)](n+1)p_{n+1}(t) + o(h). \end{aligned} \tag{2}$$

The first term on the right-hand side corresponds to no change in $N_t$ in the time interval $(t, t+h)$, the second to one birth (speciation), and the third to one death (extinction). All other events have probability $o(h)$. Subtracting $p_n(t)$ from both sides, dividing by $h$ and letting $h \to 0$, yields the differential-difference equation

$$\frac{d}{dt}p_n(t) = -(\lambda+\mu)np_n(t) + \lambda(n-1)p_{n-1}(t) + \mu(n+1)p_{n+1}(t), \tag{3}$$

with initial condition

$$p_n(0) = 1 \text{ if } n = 1; \quad p_n(0) = 0 \text{ otherwise.}$$

Now let

$$\phi(z;t) = \sum_{n=0}^{\infty} p_n(t)z^n \tag{4}$$

be the generating function for $N_t$. Multiplying both sides of (3) by $z^n$ and summing over $n = 0, \ldots, \infty$ yields the partial differential equation

$$\phi_t = (\lambda z - \mu)(z-1)\phi_z, \tag{5}$$

with initial condition

$$\phi(z,0) = z \tag{6}$$

This can readily be solved by the method of characteristics (see *e.g.* Bailey, 1964) to yield

$$\phi(z,t) = \begin{cases} \dfrac{\mu(1-z) - (\mu - \lambda z)\exp[-t(\lambda-\mu)]}{\lambda(1-z) - (\mu - \lambda z)\exp[-t(\lambda-\mu)]} & \text{if } \lambda \neq \mu, \\ 1 - (1-z)/[1 + \lambda t(1-z)]^{-1} & \text{if } \lambda = \mu, \end{cases} \tag{7}$$

From this the well-known formulas for the probability mass function (p.m.f) of $N(t)$ can be derived:

$$p_0(t) \;=\; \frac{\mu - \mu e^{-t(\lambda-\mu)}}{\lambda - \mu e^{-t(\lambda-\mu)}}; \tag{8}$$

$$p_n(t) = \;=\; \frac{(\lambda-\mu)^2 e^{-t(\lambda-\mu)}}{[\lambda - \mu e^{-t(\lambda-\mu)}]^2} \left\{ \frac{\lambda - \lambda e^{-t(\lambda-\mu)}}{\lambda - \mu e^{-t(\lambda-\mu)}} \right\}^{n-1}, \quad n \geq 1, \tag{9}$$

for $\lambda = \mu$ and

$$p_0(t) = \frac{\lambda t}{1 + \lambda t}; \qquad p_n(t) = \frac{(\lambda t)^{n-1}}{(1+\lambda t)^{n+1}}, \quad n \geq 1. \tag{10}$$

in the case $\lambda = \mu$.

Since different genera in a family will have originated at different times, in order to obtain the p.m.f (or generating function) of the unconditional distribution of genus size $\bar{N}$ say, $\{p_n(t)\}$ (or $\phi(z,t)$) must be integrated with respect to the distribution of $t$ over genera. Following Yule, it seems reasonable to assume that new genera originated from old by a process analogous to that of speciation. Under this assumption the time since origin of any genus will be exponentially distributed with parameter $\rho$ (where $\rho h + o(h)$ is the probability of any genus splitting in time $(t, t+h)$). It follows that the p.m.f of the unconditional (marginal) distribution of genus size, $\bar{N}$, is

$$q_n = \Pr(\bar{N} = n) = \int_0^\infty p_n(t)\rho e^{-\rho t} dt \quad \text{for } n = 0, 1, \ldots \tag{11}$$

and its generating function

$$\bar{\phi}(z) = \int_0^\infty \phi(z,t)\rho e^{-\rho t} dt. \tag{12}$$

Neither of the integrals in (11) or (12) have simple closed-form expressions. However by returning to the partial differential equation (5), multiplying throughout by $\rho e^{-\rho t}$ and integrating with respect to $t$ between 0 and $\infty$, one arrives at the following ordinary differential equation for $\bar{\phi}(z)$:

$$(\lambda z - \mu)(z-1)\frac{d\bar{\phi}}{dz} - \rho\bar{\phi}(z) = -\rho z. \tag{13}$$

This can be solved using the integrating factor

$$\exp\left(-\int^z \frac{\rho dz}{(\lambda z - 1)(z-1)}\right) \propto \begin{cases} \left|\dfrac{z-1}{\lambda z - \mu}\right|^{\frac{\rho}{|\lambda-\mu|}} & \text{if } \lambda \neq \mu, \\ \exp\left(\frac{-\rho}{\lambda(z-1)}\right) & \text{if } \lambda = \mu \end{cases} \tag{14}$$

and integrating between $z$ and $\frac{\mu}{\lambda}$ for $\lambda > \mu$ and between $z$ and 1 for $\lambda \geq \mu$. The result is

$$\bar{\phi}(z) = \begin{cases} 1 - \dfrac{\rho}{\lambda}\mathrm{L}\left(\dfrac{\mu - \lambda z}{\lambda - \lambda z}, \dfrac{\rho}{\lambda - \mu}\right) & \text{if } \lambda > \mu, \\ 1 - \frac{\rho}{\lambda}\exp\left(-\frac{\rho}{\lambda(z-1)}\right)\mathrm{E}_1\left(-\frac{\rho}{\lambda(z-1)}\right) & \text{if } \lambda = \mu, \\ \frac{\mu}{\lambda} - \frac{\rho}{\lambda}\mathrm{L}\left(\frac{\lambda-\lambda z}{\mu-\lambda z}, \frac{\rho}{\mu-\lambda}\right) & \text{if } \lambda < \mu. \end{cases} \tag{15}$$

where L is Lerch's phi function

$$\mathrm{L}(x,\theta)=\sum_{n=0}^{\infty}\frac{x^n}{(n+\theta)} \tag{16}$$

and $\mathrm{E}_1$ is the exponential integral function

$$\mathrm{E}_1(x)=\int_1^{\infty}\frac{e^{-xt}}{t}dt. \tag{17}$$

In principle, the p.m.f $\{q_n\}$ of genus size can be obtained by expanding the generating function (15) as a Taylor series around $z=0$. However this is not practical beyond a few terms. The p.m.f can be computed numerically by numerical integration of (11). However this can be slow and unreliable for larger values of $n$, and when $|\lambda-\mu|$ is small. An alternative method described in Section 5 relies on the asymptotic behaviour of $q_n$ which is described in the next section.

# 3 Properties of the p.m.f. of genus size.

Although there are three parameters in the model, a glance at (7) - (12) will confirm that in determining the p.m.f $\{q_n\}_{n=0,1,\ldots}$ they are not independent and that the p.m.f can be expressed in terms of any two ratios of the three parameters[2]. We shall use the two ratios $\tilde{\lambda}=\lambda/\rho$ and $\tilde{\mu}=\mu/\rho$.

Yule (1924) observed that for his model ($\mu=0$), the plot of the p.m.f (the Yule distribution) on logarithmic axes was almost linear. How is this property

[2] Alternatively we could set one of the parameters to unity, which effectively means defining the unit of time (*e.g.* setting $\mu=1$ is equivalent to defining the unit of time as the expected time for a species to exist before becoming extinct).

affected by the inclusion of the possibility of extinctions in the model? Fig.1 gives logarithmic plots with $\tilde{\lambda} = 10$ and $\tilde{\mu} = 10, 9, 8$ and $0$[3]. It can be seen that the inclusion of the possibility of extinctions introduces some curvature into the plots. However if $\tilde{\mu}$ is small the curvature is negligible. It is only as $\tilde{\mu}$ approaches $\tilde{\lambda}$ that the curvature becomes significant. Changing the value of $\tilde{\lambda}$ seems to have little effect on this conclusion.

The expected size of a genus is

$$\mathrm{E}(\bar{N}) = \begin{cases} \rho/(\rho + \mu - \lambda), & \text{if } \lambda < \mu + \rho; \\ \infty, & \text{if } \lambda \geq \mu + \rho. \end{cases} \tag{18}$$

This can be established by evaluating the first derivative of $\bar{\phi}$ at $s = 1$ or more easily by evaluating the integral $\int_0^\infty e^{(\lambda-\mu)t} \rho e^{-\rho t} dt$. The corresponding expected value for a genus, given that it is non-empty (*i.e.* of a live genus) can be obtained by dividing $\mathrm{E}(\bar{N})$ by $1 - q_0$.

The asymptotic behaviour of $q_n$ in the three cases (a) $\lambda > \mu$; (b) $\lambda < \mu$ and (c) $\lambda = \mu$ is examined in the Appendix, where it is shown that:

(a) for $\lambda > \mu$

$$q_n \sim c_1 n^{-(\rho/(\lambda-\mu)+1)}; \tag{19}$$

(b) for $\lambda < \mu$

$$q_n \sim c_2 \left(\frac{\lambda}{\mu}\right)^n n^{-(\rho/(\mu-\lambda)+1)}; \tag{20}$$

and (c) for $\lambda = \mu$

$$q_{n+1} \sim \frac{\pi^{1/2}(\rho/\lambda)^{5/4}}{n^{3/4}} \exp[-2(\rho/\lambda)^{1/2} n^{1/2}]. \tag{21}$$

[3] For the method used for computing the probabilities in the p.m.f, see Sec.5

where $c_1$ and $c_2$ are constants. Thus for $\lambda > \mu$ the p.m.f of genus size exhibits power-law behaviour asymptotically; while when $\lambda < \mu$ asymptotically the p.m.f decays faster than a power law. In the threshold case $\lambda = \mu$ the decay is also faster than a power law, being of a stretched exponential form.

Note that for a live (non-empty) genus, the p.m.f is obtained by dividing the $q_n$ by $1 - q_0$. The constants in the formulas for the asymptotic behaviour of the p.m.f require similar adjustment. The above asymptotic formulas are useful in computing the p.m.f (see Sec. 5).

# 4 Time since establishment of a genus.

Under the assumptions in Section 2, it is possible to derive the distribution of the time since establishment, $T$, of a genus currently containing $n > 0$ species. From Bayes' theorem, the density of the time since establishment is

$$f(t|n) \propto \rho e^{-\rho t} \frac{p_n(t)}{1 - p_0(t)} \tag{22}$$

with the constant of proportionality being $(1 - q_0)/q_n$. Thus from (8) - (10) for $\lambda \neq \mu$

$$f(t|n) \propto \frac{e^{-(\lambda-\mu+\rho)t}}{\lambda - \mu e^{-(\lambda-\mu)t}} \left( \frac{\lambda - \lambda e^{-(\lambda-\mu)t}}{\lambda - \lambda e^{-(\lambda-\mu)t}} \right)^{n-1} \qquad t > 0 \tag{23}$$

while for $\lambda = \mu$ it is

$$f(t|n) \propto e^{-\rho t} \frac{(\lambda t)^{n-1}}{(1 + \lambda t)^n} \qquad t > 0 \tag{24}$$

Fig. 2 shows the density of $T$ for genera containing respectively, 1,10,100 and 453 species, using estimated parameter values for N. American vascular plants (see Section 6). The time unit is the expected time for a genus to give rise to a new genus ($1/\rho$).

# 5 Computation of the p.m.f. of genus size.

From the differential equation (13) for the generating function $\bar{\phi}(z)$ it is possible to obtain a recurrence relation for the probabilities $\{q_n\}$ in the p.m.f of $\bar{N}$. This is done simply by identifying the coefficient of $z^n$ on each side of the equation. This yields

$$(n-1)\lambda q_{n-1} - [n(\lambda+\mu)+\rho]q_n + (n+1)\mu q_{n+1} = 0, \quad \text{and} \tag{25}$$

$$-(\lambda+\mu+\rho)q_1 + 2\mu q_2 = -\rho \tag{26}$$

$$-\rho q_0 + \mu q_1 = 0. \tag{27}$$

While it is possible to iterate this recursion forwards from $q_0$, unless one uses very high precision arithmetic, round-off error soon becomes a problem. A better alternative is to iterate it backwards from a large value $\nu$ say of $n$, using the asymptotic results for $q_\nu$ and $q_{\nu+1}$. In detail the method works as follows: first write $q_n = cX_n$ and suppose that for large $\nu$, $q_\nu$ has the asymptotic form derived Sec.3. This means, in the case $\lambda > \mu$ that $X_\nu = \nu^{-(\rho/\lambda+1)}$; and in the case $\lambda < \mu$ that $X_\nu = (\lambda/\mu)^\nu \nu^{-(\rho/\lambda+1)}$ *etc.* The recursion (25) holds for $X_n$ as well as for $q_n$, and the next step involves iterating (25) backwards (starting with $X_{\nu+1}$ and $X_\nu$) to obtain $X_2$. But $X_2$ can also be expressed in terms of $q_0$ (using (26) and (27)) as $(\rho(\mu+(\lambda+\mu+\rho)q_0)/2c\mu^2$ and $q_0$ can be evaluated as $\bar{\phi}(0)$, using (15) (or, more slowly, by numerically evaluating the integral (11)). Thus the constant $c$ can be evaluated as

$$c = \frac{\rho(\mu + (\lambda + \mu + \rho)q_0)}{2\mu^2 X_2}. \tag{28}$$

The probabilities $q_n$ are then determined as $q_n = cX_n$, and the probability that a non-empty genus is of size $n$ is given by $cX_n/(1 - q_0)$.

# 6 Maximum likelihood estimation.

We consider data of the type given in Yule (1924), Burlando (1990, 1993) where the numbers of species in each of many genera in a family are recorded. Thus suppose $f_i$ genera containing $i$ species ($i = 1, \ldots, N$, say) are observed. Assuming independence the log-likelihood for such data is

$$\ell = \sum_{i=1}^{N} f_i \log q_i - \log(1 - q_0) \sum_{i=1}^{N} f_i \tag{29}$$

the latter term being present because only living (non-empty) genera are observed. The parameters $\lambda, \mu$ and $\rho$ enter the log-likelihood *via* the $q_i$, but as noted before only two ratios the three are needed to determine the likelihood completely. As before we shall use $\tilde{\lambda} = \lambda/\rho$ and $\tilde{\mu} = \mu/\rho$. Numerical computation of the log-likelihood, for particular values of $\tilde{\lambda}$ and $\tilde{\mu}$ involves first calculating the $q_n$ by the method outlined in Sec.5. To fit the model by maximum likelihood (ML) the log-likelihood must be maximized (numerically) with respect to $\tilde{\lambda}$ and $\tilde{\mu}$.

We illustrate with using two datasets (a) N. American vascular plants (Qian and Ricklefs, 2000) which has 1829 genera, the largest having 453 species; (b) the data on snakes used by Yule (1924) with 293 genera, the largest having 97 species.

### *North American vascular plants.*

Fig. 3 shows a plot of the frequencies $f_i$ of genera of size $i$ against $i$ (logarithmic axes) both ungrouped (crosses) and grouped for the rarer large genus sizes (boxes). Also shown is the fitted distribution (curved line)

$$\left(\sum_{i=1}^{N} f_i\right) \frac{\hat{q}_i}{1-\hat{q}_0}, \qquad i = 1, 2, \ldots, N \tag{30}$$

where the $\hat{q}_i$ are the probabilities in the p.m.f evaluated using the M.L estimates of parameter ratios $\tilde{\lambda}$ and $\tilde{\mu}$). The log-likelihood surface has a narrow ridge close to the line $\tilde{\lambda} = \tilde{\mu}$, and hence the ML estimates are highly correlated. Because of this fact it is better to consider a transformation of the parameters. Thus consider

$$\theta_1 = \tilde{\lambda} + \tilde{\mu} \qquad\qquad \theta_2 = \tilde{\lambda} - \tilde{\mu}. \tag{31}$$

Fig 4 shows a contour plot of the log-likelihood in $(\theta_1, \theta_2)$-space, with contours corresponding to 1%, 5% and 10% likelihood regions, or approximate 99%, 95% and 90% confidence regions[4]. It can be seen that with a high degree of confidence we can conclude that the difference $\tilde{\lambda} - \tilde{\mu}$ lies between about 0.2 and 0.8; and that the average of $\tilde{\lambda}$ and $\tilde{\mu}$ is between 5 and 15. The ML estimates of $\tilde{\lambda}$ and $\tilde{\mu}$ are respectively 9.008 and 8.478. Although the estimates

[4] For a discussion of likelihood intervals and regions see *e.g.* Kalbfleisch, 1985. Their interpretation as approximate confidence regions is based on asymptotic likelihood ratio theory.

are close there is strong evidence to indicate that they are not equal - the line $\theta_2 = 0$ lies well outside of the 99% confidence region[5]. Thus we can conclude that there is a small but important difference between the speciation and extinction rates.

The chi-square goodness-of-fit statistic using the 40 groups (as shown by boxes in Fig. 3) is 74.32 which is highly significant. However bearing in mind the fact that there are 1829 genera, this is hardly surprising. Given the way that the frequencies of genera of sizes in the range 80-120 occur, it seems likely that any parametrically parsimonious distribution, would exhibit a similarly large chi-square statistic.

The fitted distribution in Fig. 3 exhibits some curvature even though asymptotically it is linear (power-law behaviour prevails). Yule (1924) attributed the departures from linearity that he observed to finite time effects. An alternative explanation is the presence of ongoing individual extinctions.

## *Snakes.*

Fig. 5 shows observed and fitted frequencies of genera over the range of genus sizes. The M.L. estimates were 10.007 for $\tilde{\lambda}$ and 9.991 for $\tilde{\mu}$, which are very close and close to those for N. American vascular plants. In fact there is no evidence of a significant difference between $\tilde{\lambda}$ and $\tilde{\mu}$ (P= 0.93). If they are assumed equal the ML estimate of their common value is 9.831 and a 95% confidence interval for it is $(7.22, 13.69)$. Thus although the individual

[5] In fact the likelihood ratio (Royall, 1997) for the hypothesis $\tilde{\lambda} = \tilde{\mu}$ versus the alternative of inequality is less than $10^{-5}$.

estimates for speciation and extinction rates are similar for vascular plants and snakes, in the former case, unlike in the latter, there is strong evidence that the two rates differ. Possible reasons for this are the smaller sample size for the snake data, and also shortcomings in the data which dates from 1893 (see Yule,1924).

The chi-square goodness-of-fit statistic using 17 groups is 21.57 (P = 0.09). The fitted distribution (Fig.5) for snakes exhibits a similar degree of curvature to that for vascular plants.

## 7 Concluding remarks.

In this article we have investigated some of the properties of genus size distribution arising from a 'null model' in which speciations and extinctions occur independently and at random, and in which new genera are created in an analogous way to species. The resulting distribution, which depends on two independent parameters, is shown to be capable of exhibiting behaviour qualitatively similar to observed size distributions. It always has its mode at one (the monospecific genus has the highest probability of occurring), and when the probability of a genus of size $n$ is plotted against $n$ on logarithmic axes, the points initially lie close to a straight line, and if any curvature is present it is always in a downward direction.

Yule (1924) considered a similar model, but without the possibility of extinctions. In its simplest form Yule's model produced power-law behaviour - linear logarithmic plots of probability *vs.* size, but Yule observed that in empirical logarithmic plots there were often departures from linearity at higher abundance levels. He attributed this as being due to finite time effects. This paper has shown that another possible explanation is the presence of individual extinctions, occurring at random. In principle it would be possible to include multiple extinctions - indeed it is possible to set up a recurrence analogous to that in Section 3 in the case when the probability of $j$ extinctions occurring in the infinitesimal interval $(t, t + dt)$ in a genus of size $n$ is $\mu_{n,j} dt$. Such a model could lead to episodic mass extinctions, as has been observed in the fossil record (Raup, 1985, 1991). However there are difficulties in analysing such a model and furthermore it is not obvious how such multiple extinction probabilities should be specified.

When the abundance distribution derived from the model was fitted to observed species abundance data, it provided a good fit, and furthermore provided rather similar estimates for vascular plants and for snakes. In both cases the estimates of extinction and speciation rates were close. It is possible that the fossil record could provide data on extinctions which could be used to test the model.

A related problem, dealt with in a forthcoming paper (Reed and Hughes, 2001), concerns the size distribution of extinct fossil genera (or higher taxa). Under similar assumptions concerning speciations and individual extinctions, and including the possibility of catastrophic extinctions in which all species in a genus (or sub-taxa in a taxon) are destroyed, to determine the size distribution one needs to determine the distribution of the number of species which *have ever existed* in the genus, before it became extinct. Mathematically this is a more difficult problem, but expressions for the generating function of the distribution can be obtained. In addition it is possible to determine the distribution of the lifetime of a genus. It should be possible to fit this model to empirical fossil abundance distributions and to taxon lifetime data.

### Acknowledgements.

We gratefully acknowledge the contributions of Fangliang He, for his comments and encouragement and for providing the data collected by Hong Qian; and of Louis Doray, for pointing out several useful references.

## References.

Burlando, B. (1990). The fractal dimension of taxonomic systems. *J. Theor. Biol.* 146:99-114.

Burlando, B. (1993). The fractal geometry of evolution. *J. Theor. Biol.* 163:161-172.

Chu, J. and C Adami (1999) A simple explanation for taxon abundance patterns. *Proc. Nat. Acad. Sci.* 96:15017-15019.

Corbet, A. S. (1942) The distribution of butterflies in the Malaysian Peninsula. *Proc. R. Entomol. Soc. Lond. A* 16:101-116.

Kalbfliesch, J. G. (1985) *Probability and Statistical Inference. Vol. 2: Statistical Inference.* Springer-Verlag, New York.

Kemp, A. W. (1995) Splitters, lumpers and species per genus. *Math. Scientist* 20:107-118.

Qian, H. and Ricklefs, R. E (2000). Large-scale processes and the Asian bias in species diversity of temperate plants. *Nature* 407:180-782.

Raup, D. M. (1985) Mathematical models of cladogenesis. *Paleobiology* 11:42-52.

Raup, D. M. (1991) *Bad Genes or Bad Luck?* W. W Norton & Co., New York.

Reed, W. J. and B. D. Hughes (2001). A model for the size distribution of extinct taxa. In preparation.

Royall, R (1997). *Statistical Evidence. A Likelihood Paradigm.* Chapman & Hall, London.

Yule, G. U. (1924) A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis. F.R.S. *R. Soc. Lond., Philos. Trans. (B)* 213:21-87.

# Appendix. Asymptotic behaviour of the genus size distribution.

The asymptotic behaviour of the p.m.f of genus size $q_n$, as $n \to \infty$ is examined in the three cases given in Sec.3.

## Case (a) $\lambda > \mu$

Consider

$$q_{n+1} = \int_0^\infty \frac{\rho e^{-\rho t}(\lambda-\mu)^2 e^{-t(\lambda-\mu)}}{[\lambda-\mu e^{-t(\lambda-\mu)}]^2} \left\{ \frac{\lambda - \lambda e^{-t(\lambda-\mu)}}{\lambda - \mu e^{-t(\lambda-\mu)}} \right\}^n dt. \tag{32}$$

The factor in braces is strictly increasing in $t$ from zero at $t = 0$, approaching unity as $t \to \infty$. Effecting a change of variable from $t$ to $\tau$ with

$$t = (\lambda - \mu)^{-1} \log[(n(1 - \mu/\lambda)/\tau], \tag{33}$$

we have as as $n \to \infty$,

$$\left\{\frac{\lambda - \lambda e^{-t(\lambda-\mu)}}{\lambda - \mu e^{-t(\lambda-\mu)}}\right\}^n \sim \left\{1 - \frac{\tau}{n}\right\}^n \to e^{-\tau}. \tag{34}$$

so that in (32)

$$q_{n+1} \sim \frac{\rho}{\lambda}\Big(1 - \frac{\mu}{\lambda}\Big)^{-\rho/(\lambda-\mu)} n^{-1-\rho/(\lambda-\mu)} \int_0^\infty e^{-\tau}\tau^{\rho/(\lambda-\mu)} d\tau. \tag{35}$$

The integral can be evaluated as a gamma function. The important conclusion is that for large $n$, the p.m.f exhibits power-law behaviour *i.e.*

$$q_n \sim c_1 n^{-(\rho/(\lambda-\mu)+1)}. \tag{36}$$

where

$$c_1 = \frac{\rho}{\lambda}\Big(1 - \frac{\mu}{\lambda}\Big)^{-\rho/(\lambda-\mu)} \Gamma(1 + \rho/(\lambda - \mu)) \tag{37}$$

The reason for the divergence of $\mathrm{E}(\bar{N})$ when $\lambda > \mu + \rho$ is now apparent since in this case the exponent of $n$ in (36) lies between -1 and -2.

## Case (b) $\lambda < \mu$

Using the change of variable

$$t = (\mu - \lambda)^{-1} \log[(n(1 - \lambda/\mu)/\tau], \tag{38}$$

we have as as

$$\left\{\frac{\lambda - \lambda e^{-t(\lambda-\mu)}}{\lambda - \mu e^{-t(\lambda-\mu)}}\right\}^n \sim \left(\frac{\lambda}{\mu}\right)^n e^{-\tau}. \tag{39}$$

which in (32) yields the result that as $n \to \infty$,

$$q_n \sim c_2 \left(\frac{\lambda}{\mu}\right)^n n^{-(\rho/(\mu-\lambda)+1)} \tag{40}$$

where

$$c_2 = \frac{\rho}{\lambda}\left(1 - \frac{\lambda}{\mu}\right)^{-\rho/(\mu-\lambda)}\Gamma(1+\rho/(\mu-\lambda)) \tag{41}$$

Thus when the individual extinction rate exceeds the speciation rate the p.m.f. decays faster than a power law.

## Case (c) $\lambda = \mu$

In this case, we know that $\mathrm{E}(N_t) = 1$ for all $t$, so that $\mathrm{E}(\bar{N}) = 1$, and we may anticipate that the p.m.f. of $\bar{N}$ is reasonably rapidly decaying, though its dominant form is not obvious. Again consider

$$q_{n+1} = \int_0^\infty \frac{\rho e^{-\rho t}}{(1+\lambda t)^2}\left\{\frac{\lambda t}{1+\lambda t}\right\}^n dt. \tag{42}$$

Inspecting the relative sizes of $e^{-\rho t}$ and the term in braces, suggests a change of variable from $t$ to $\tau$ where $t = n^{1/2}\tau/\lambda$, so

$$q_{n+1} = \frac{\rho n^{1/2}}{\lambda}\int_0^\infty \frac{e^{-(\rho/\lambda)n^{1/2}\tau}}{(1+n^{1/2}\tau)^2}\left\{1 + \frac{1}{n^{1/2}\tau}\right\}^{-n} d\tau. \tag{43}$$

Noting that as $n \to \infty$, $\left\{1 + \frac{1}{n^{1/2}\tau}\right\}^{-n^{1/2}} \to e^{-\tau^{-1}}$, we see that

$$q_{n+1} \sim \frac{\rho n^{1/2}}{\lambda}\int_0^\infty \frac{\exp\left(-n^{1/2}p(\tau)\right) d\tau}{(1+n^{1/2}\tau)^2}, \tag{44}$$

where

$$p(\tau) = (\rho/\lambda)\tau + \tau^{-1}. \tag{45}$$

The integral is now suited to the application of the method of Laplace. The mild $n$-dependence of the denominator presents no obstacles, but we refrain from writing out the details, simply noting that $p(\tau)$ is minimized at $\tau = (\lambda/\rho)^{1/2}$, and that in the neighbourhood of this point,

$$p(\tau) = 2(\rho/\lambda)^{1/2} + (\rho/\lambda)^{3/2}[\tau - (\lambda/\rho)^{1/2}]^2 + \cdots. \tag{46}$$

Hence

$$q_{n+1} \sim \frac{\rho n^{1/2}}{\lambda} \int_{-\infty}^{\infty} \frac{\exp(-2(\rho/\lambda)^{1/2}n^{1/2} - (\rho/\lambda)^{3/2}n^{1/2}[\tau - (\lambda/\rho)^{1/2}]^2)d\tau}{(1 + n^{1/2}(\lambda/\rho)^{1/2})^2}, \tag{47}$$

and since $\int_{-\infty}^{\infty} e^{-x^2}dx = \sqrt{\pi}$, the integral is now exactly evaluable, giving

$$q_{n+1} \sim \frac{\pi^{1/2}(\rho/\lambda)^{5/4}}{n^{3/4}} \exp[-2(\rho/\lambda)^{1/2}n^{1/2}]. \tag{48}$$

and we see that the p.m.f decays faster than a power law.

Note that for a live (non-empty) genus, the p.m.f is obtained by dividing the $q_n$ by $1 - q_0$. The constants in the formulas for the asymptotic behaviour of the p.m.f require similar adjustment.

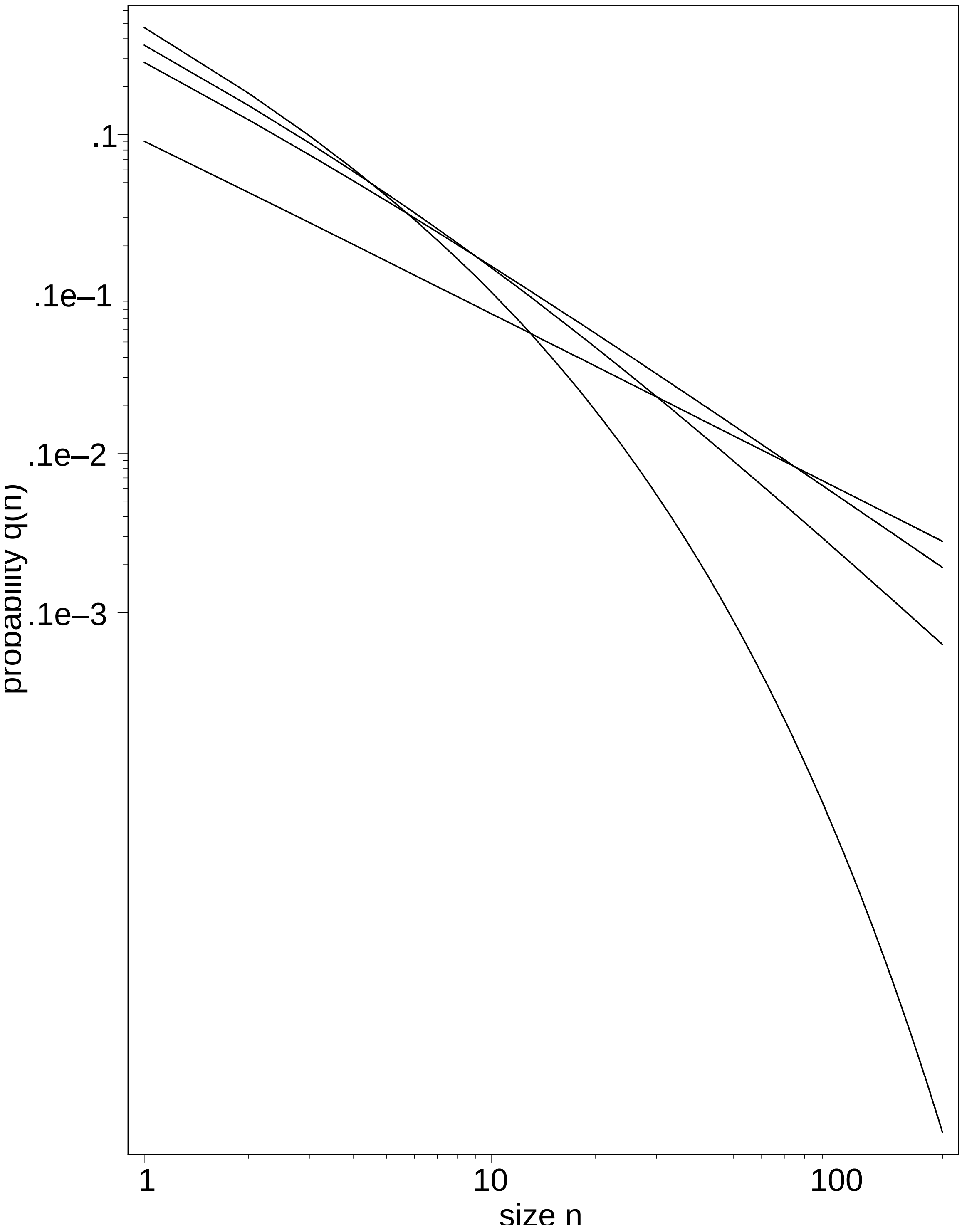

Figure 1: Logarithmic plots of the probability mass function of genus size. The parameter $\tilde{\lambda}$ is set at 10, and $\tilde{\mu}$ takes values (downward from top on left-hand side) 10, 9,8 and 0. For presentation purpose the individual points have been joined with a smooth curve.

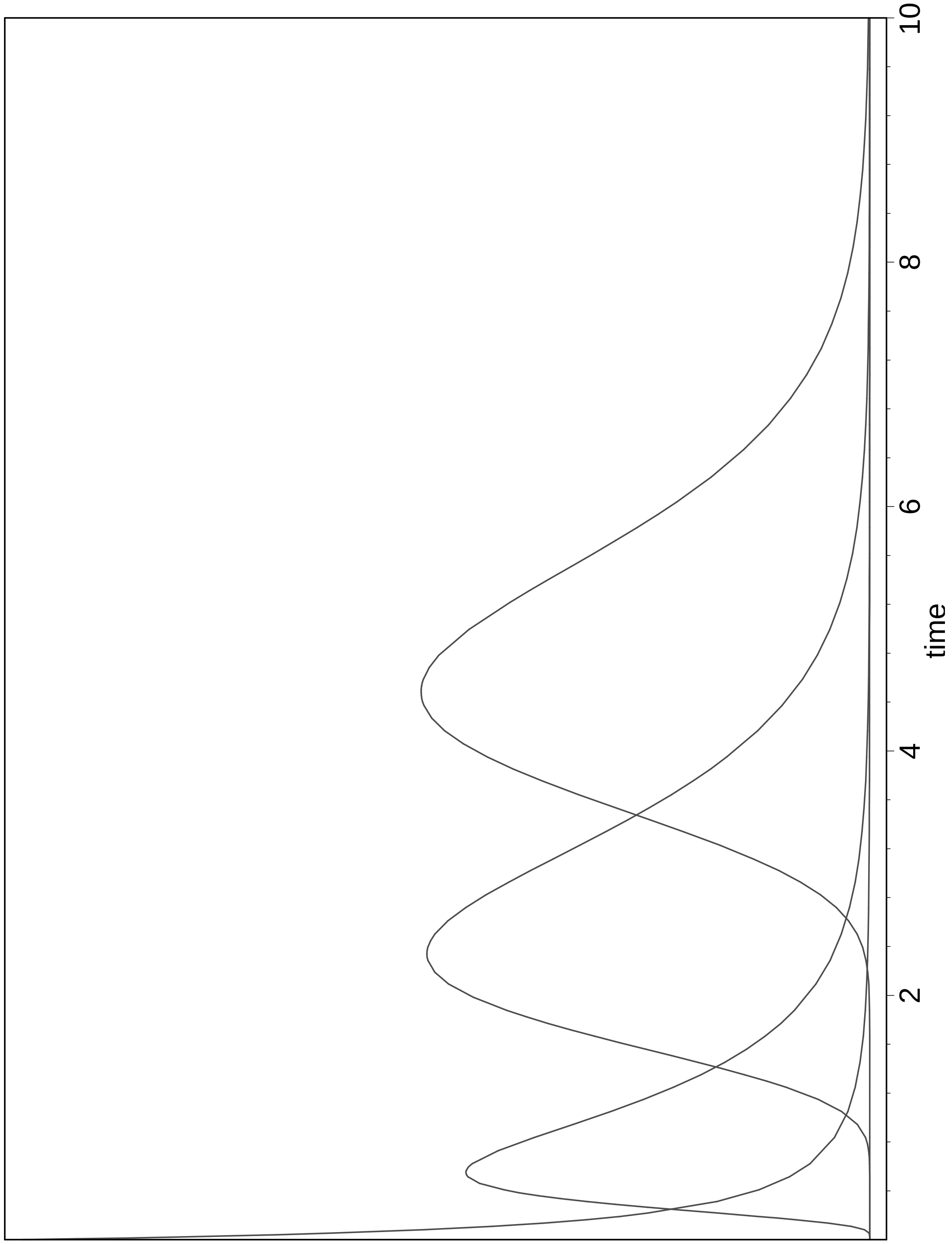

Figure 2: The probability density function of the time since establishment of genera containing (from the left) 1, 10, 100 and 453 species. The parameter values used are the maximum likelihood estimates for N. American vascular plants. The time unit is the expected time for a genus to give rise to a new genus ($1/\rho$).

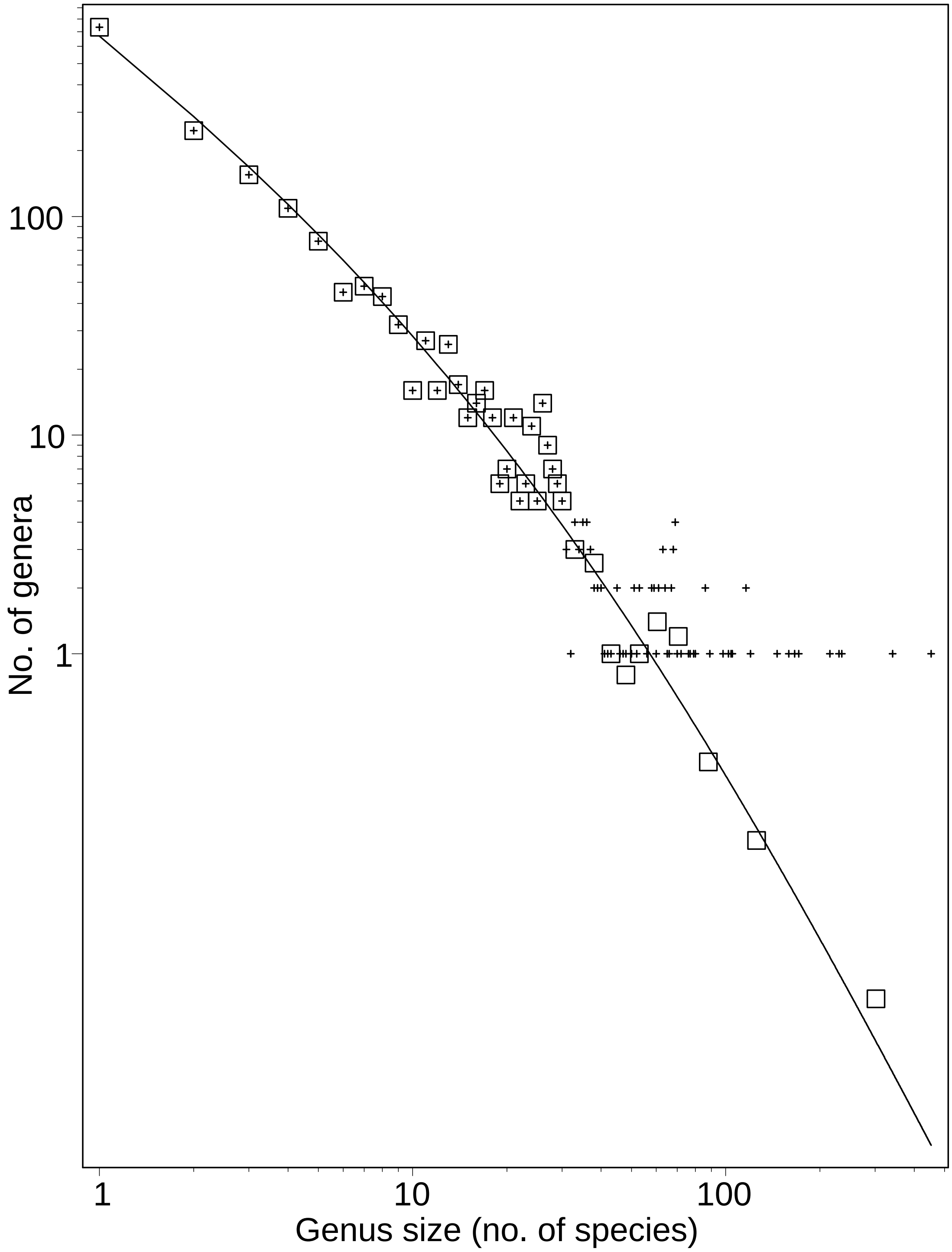

Figure 3: Logarithmic frequency plot of the observed sizes of 1829 genera of North American vascular plants (crosses). Also shown is a similar plot with the larger less frequent genera grouped (boxes), and the fitted distribution (with points joined by a smooth curve for display purposes).

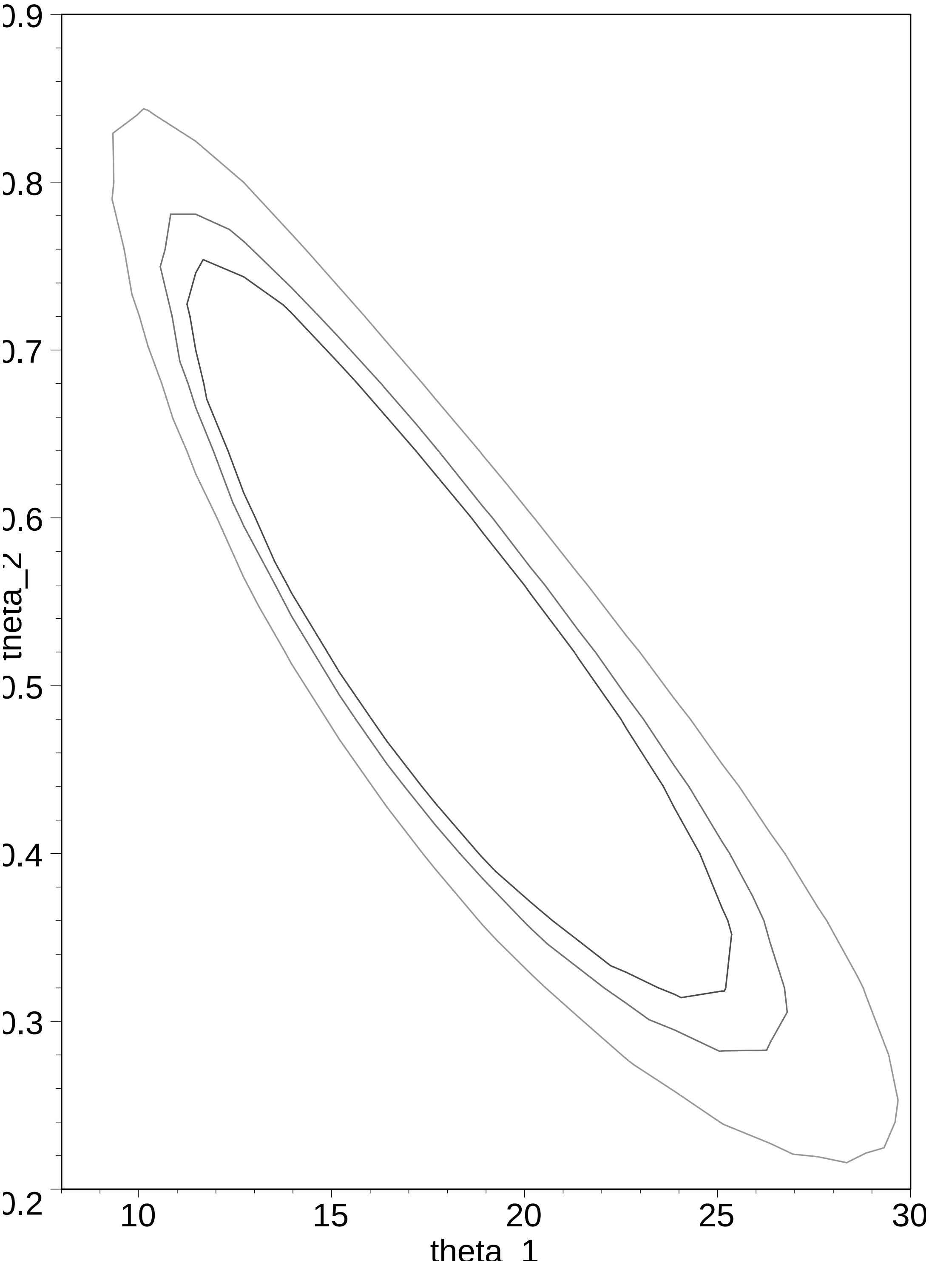

Figure 4: A plot of the contours of the log-likelihood for N. American vascular plants, expressed in terms of the parameters $\theta_1 = \tilde{\lambda} + \tilde{\mu}$ and $\theta_2 = \tilde{\lambda} - \tilde{\mu}$. The contours correspond to (working outwards) 10%, 5% and 1% likelihood regions, or approximate 90%, 95% and 99% confidence regions.

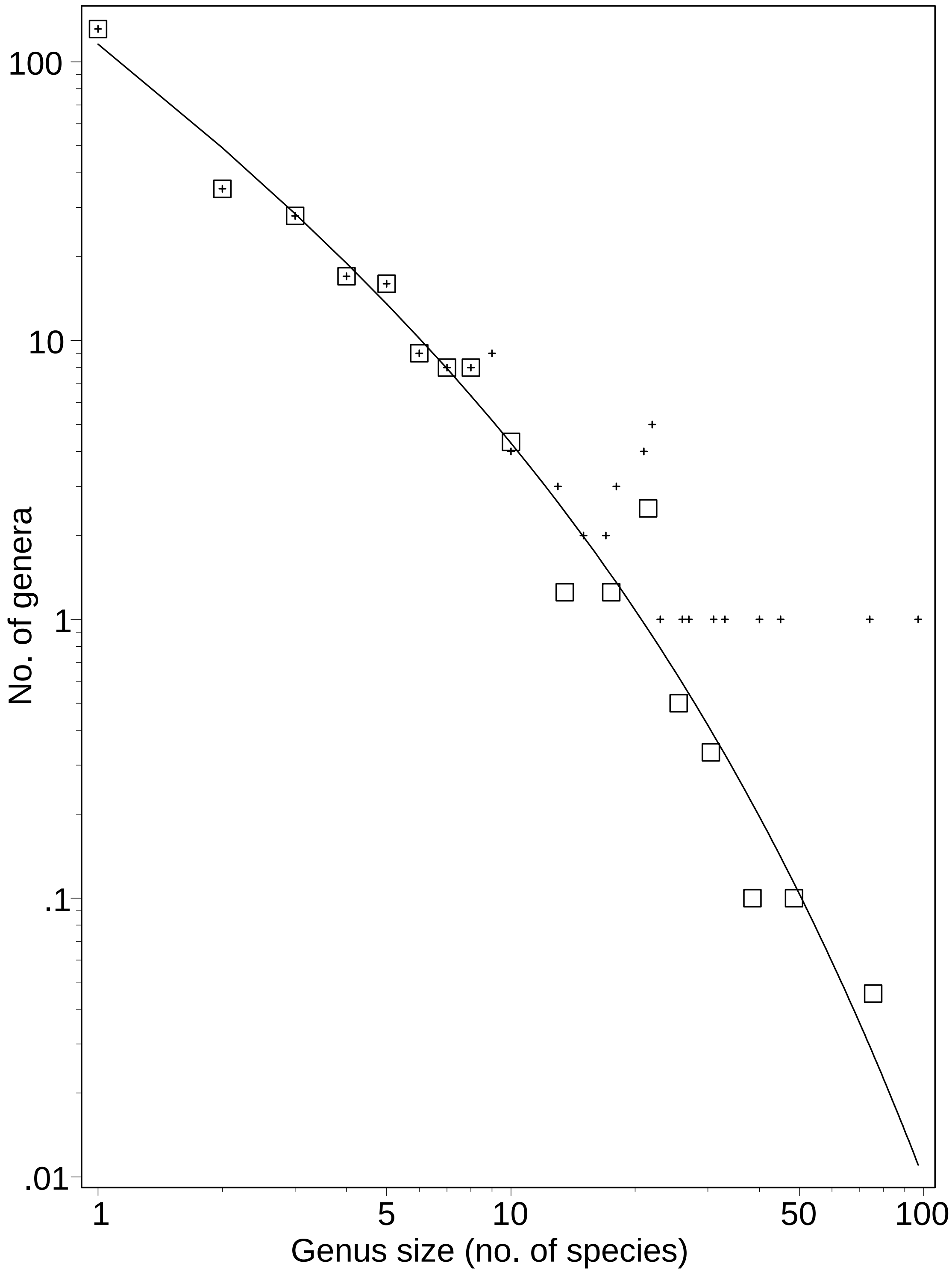

Figure 5: Logarithmic frequency plot of the observed sizes of 293 genera of snakes (crosses). Also shown is a similar plot with the larger less frequent genera grouped (boxes), and the fitted distribution (with points joined by a smooth curve for display purposes).