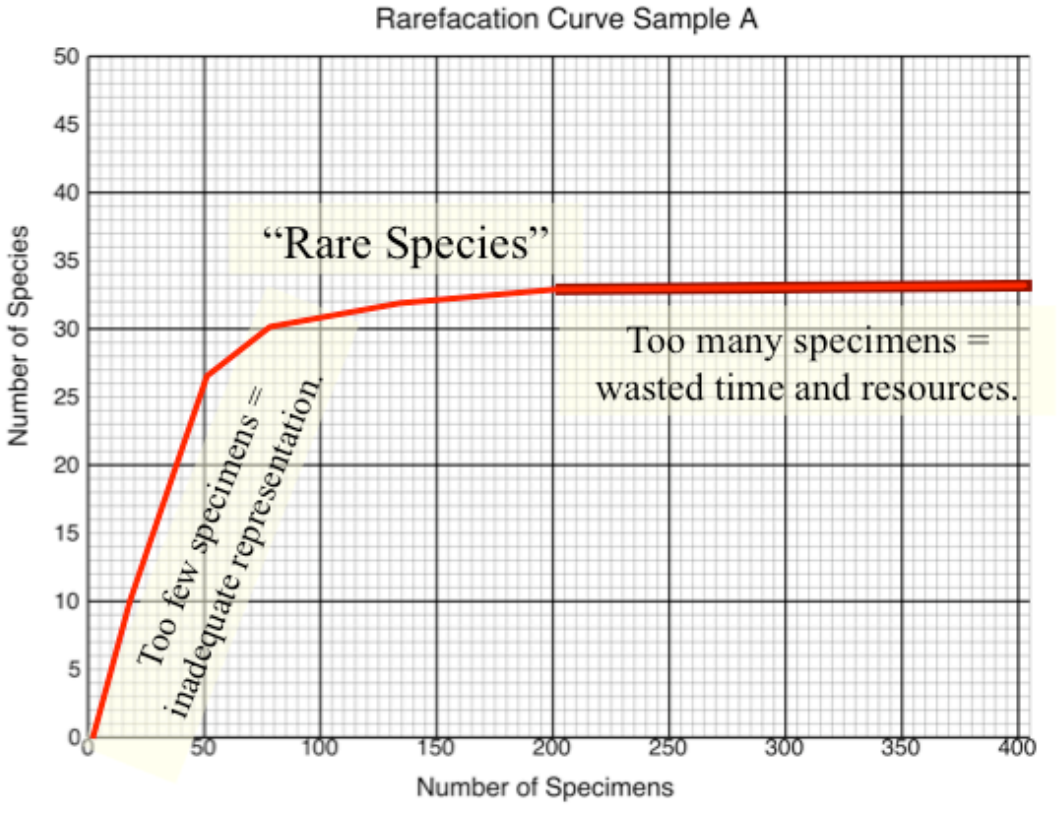
Principles of Paleontology: Rarefaction

".. how much is enough"

When presented with a collection of hundreds or thousands (or even hundred of thousands) of specimens, how many individual specimens must you identify before you are confident that you have seen *all* of the species represented in the pile (and is the pile itself even large enough to represent all of the species from the collection it represents?) These questions can be addressed with methods known as rarefaction.

In general, rarefaction is defined as "The process of becoming light or less dense." The usage here is to minimize the number of observations needed (become less dense) to determine the number of species present. That is, there is no need to identify 1,000 specimens, if you can be confident after the first 100 that there are only two species present. Of potentially even more value, the methods allow you to recognize the situation where you have identified 200 specimens, but are still finding new species.

In this exercise you will sub-sample a larger assemblage and at each stage plot the total number of specimens you have observed and the number of species recognized. Typically, at the completion of an exercise you will produce a curve, with three general parts. 1) steep slope – indicates the rapid and continual addition of species; 2) gentle slope – represents the addition of rarer species (note: this is a *generalization*, statistically common species could by chance just not have been collected until later) – typically though, the majority of common species will already have been encountered; 3) when all species have been identified, the curve will flatten to horizontal (note that *exceptionally* rare, or unlikely to be missed (but possible) could still appear late in the process). However, once a curve has flattened for a substantial distance, one becomes more confident that all species have been recognized.



Exercise Part I

1. Pour out Assemblage A onto the sorting grid (spread randomly and approximately evenly).

2. Start in grid "#1" lower left and select the first five specimens. (gather the closest five without regard to anything else.

3. Sort the specimens into discrete species using your own criteria. We will discuss details of grastropod morphology later, just use all of the features you see.

4. On the Graph "Rarefaction Curve for Assemblage A," plot number of species (Y) for first five specimens (X).

5. Repeat steps 2-4 working your way five specimens at a time through the first 30 specimens (6 rounds).

Questions:

• *For the first 30 specimens, how many species have you found?* _____

• *Based on your rarefaction curve, do you expect to keep finding new species as you add data?* _____

6. Repeat steps 2-4 for another 30 specimens.

Questions:

• *For the first 60 specimens how many species have you found?* _____

• *Based on what you have seen so far on your rarefaction curve, do you expect to keep finding new species as you add data?* _____

• *For any given sample, write a description of how you would use a rarefaction cure to determine whether you had observed the all (or majority) of the diversity present.*

_____

_____

_____

_____

7. Go ahead an finish sampling and plotting the data (5 at time) just to be sure you did not miss any species (6 species, 78 individuals)

Rarefaction Part II: is it ever enough?

1. Pour out Assemblage A onto the sorting grid (spread randomly and approximately evenly).

2. Start in grid "#1" lower left and select the first ten specimens. (gather the closest five without regard to size, shape, similarity = random).

3. Sort the specimens into discrete species using your own criteria. On a second (empty) sorting grid begin placing specimens one species per cell. This will just help in identification and keeping a species count.

4. On the Graph "Rarefaction Curve for Assemblage A," plot number of species (Y) for first ten specimens (X).

5. Repeat steps 2-4 working your way 10 specimens at a time through the first 60 specimens (6 rounds).

Questions:

• *For the first **60** specimens how many species have you found?* _____

• *Based on what you have seen so far on your rarefaction curve, do you expect to keep finding new species as you add data?* _____

6. Repeat steps 2-4 for another 60 specimens (10 at a time) to a total of 120 specimens.

Questions:

• *For the first **120** specimens how many species have you found?* _____

• *Based on what you have seen so far on your rarefaction curve, do you expect to keep finding new species as you add data?* _____

7. Continue working, ten specimens at a time until you have sampled a total of **240** specimens. As you work, consider when you might pass into the "Rare Species" and "Too Many Specimens" regions of your curve.

Questions:

• *In hindsight, after collecting **240** specimens at which subsample (number of specimens) do you think you could say, "The majority of common species had been identified"?* _____

• *After 240 specimens, do you think that you have oversampled? (already established maximum diversity)* _____

7. Continue working, ten specimens at a time until you have sampled all **310** specimens.

Questions:

• There are 47 nominative species in this assemblage. How many species did you recognize? _____

• At which sample# did you reach your observed maximum number of species? _____

• If you <u>needed</u> to know <u>every</u> single macro-gastropod species present from the setting represented by this assemblage, how many more specimens from the setting should you sample?

_____

*• Approximately, how many species would you have observed in Assemblage B if you hand only sampled **78** individuals (like Assemblage A).* _____

• If you performed this exercise over again, would the curve be the same shape? _____

Consider this experiment. What if you *repeatedly* sub-sampled **78** specimens from Assemblage B and noted the number of species in each sub-sample. Pretend that you did this 100 times. Then you could calculate an average number of species you would expect in a 78 specimen subsample of Assemblage B (you could also calculate a 95% confidence limit). If you did this for each size of possible subsamples (n), you would generate a *true* rarefaction curve. That curve would allow you to compare sub-samples of varying size to each other (assuming that both represent subsamples of the same grand assemblage).

Optional Exercise Return to Assemblage A

1. Pour specimens on to sorting grid and randomly distribute.
2. Sample **4** specimens at a time – without replacement (choose each group of four <u>randomly</u>).
3. Record the number of species present in each group of four.
4. Calculate the average number of species observed in the 19 groups of four.
   (if other groups are participating, calculate the average for the class).

Expected number of species from 4 specimen subsample of Assemblage A _____

Calculating Rarefaction:

Remember! You can only calculate backwards (rarify), you can *not* predict unknown distributions. However, if you have multiple, related Assemblages with varying sample sizes, it can be useful to compare the Assemblages by "rarifying" them to a single standard.

Say, Assemblages have different number of observed species A=6 species, B=12 species and C=19 species, come from broadly similar settings, but each have different sample sizes (A=37 individuals, B=54 individuals, C=128 individuals). We could rarify B & C to the size of A and then compare.

We will not perform these calculations, but it is important for you to know that such comparisons can be made.

If you know:

N = total number specimens (310)
n = subsample size (78)
Ni = # of specimens of ith species
E(S) = estimated number of species present:

$$E(S) = \sum_{i-1}^{S} \left(1 - \left[\frac{\binom{N - N_i}{n}}{\binom{N}{n}} \right] \right)$$

Comparing Diversity among Assemblages:


(9 kinds, 1–2 each)

VS. (share 2)


(3 kinds, 1–6 each)

Two concepts: A) # of "kinds"
B) evenness

It is very difficult to combine these concepts mathematically.
We tend to concentrate on comparing one or the other.
Develop an **Index** for comparing diversity.

C = # in common = 2
A= # unique to first group = 6
B= # unique to second group = 1
W= # specimens in smaller group = 13

Simpson Coefficient = C / W x 100
emphasizes similarity
(more sensitive to changes)

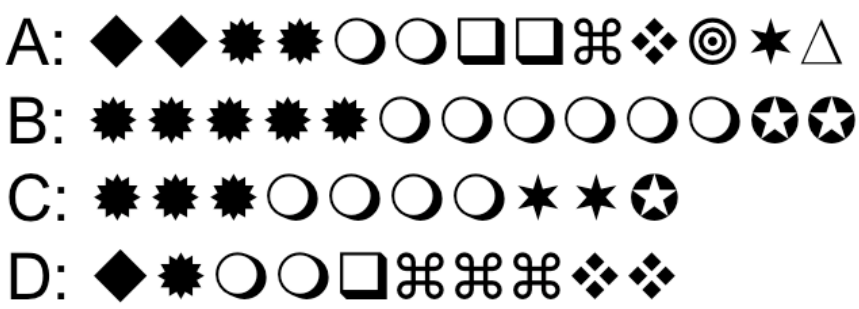
$$\frac{\text{\# of taxa in common (C) x 100}}{\text{total \# of taxa in smaller sample (W)}} = (2 \times 100)/13 = \mathbf{15.4}$$

Jaccard Coefficient = C / (A + B – C)
emphasizes differences

$$\frac{\text{\# of taxa in common (C)}}{\text{(\# exclusive to A + \# exclusive to B - \# in common)}} = 2 / (7 + 1 - 2) = \mathbf{0.33}$$

Diversity Coefficients only have meaning in the context of multiple comparisons

Calculate Jaccard Coefficients for each pair-wise comparison:



A-B _____
A-C _____
A-D _____
B-C _____
B-D _____
C-D _____

Based on Jaccard Coefficients, which two assemblages are most similar _____

which two assemblages are most dissimilar _____

For Completeness, calculate the Simpson *and* Jaccard Coefficients comparing your two gastropod assemblages.

	Assemblage A	Assemblage B
Number of individuals	W = _____	N = _____
Number of species	S_A = _____	S_B = _____
Number of shared species	C = _____	
Number of individuals unique to each	A = _____	B = _____

Simpson = _____ Jaccard = _____

30 July 30, 2009
Steven J. Hageman
Dept. of Geology
Appalachian State University
Boone, NC 28608
hagemansj@appstate.edu