Rarefaction Method

DRAFT 1/5/2016

Our data base combines taxonomic counts from 23 agencies. The number of organisms identified and counted per sample differs among agencies. Some count 100 individuals while others count more. A positive relationship exists between the probability of finding a new taxon and the number of individuals counted (Figure 1). This relationship influences important metrics such as taxa richness and diversity indices. The agency with a greater standard count will typically have higher estimates of richness and a better representation of rare taxa.

Rarefaction is often used to overcome the bias in taxa richness and diversity metrics created by different counting protocols. Rarefaction randomly selects a prescribed number of individuals (e.g., 100, 200) without replacement from a sample's larger raw count and creates a rarified sample[1]. The "vegan" package in R provides several useful rarefaction functions. The function "rarefy" provides a mean taxa richness value for the rarified sample and the error associated with that value. Mean taxa richness is the average number of taxa that can be expected in the rarified sample for a prescribed total count. Another function, "rrarefy," creates a taxonomic assemblage for a rarefied sample by randomly selecting taxa without replacement from the original sample. The rarefied assemblage can then be used as raw counts to calculate unbiased richness and diversity metrics associated with the selected total count.

Due to the random selection process, different combinations of taxa (assemblages) are usually created each time the rrarefy function is run (Table 1). Richness and diversity metric values vary with each iteration, making it difficult to score a sample consistently. The problem can be avoided by using the actual taxonomic proportions in the raw count to generate the assemblage most likely to be found in the rarified sample. When the rrarefy function is repeated many times, and the frequency of each unique taxonomic assemblage is plotted, a normalized curve is produced. The most frequently found assemblages occur in the middle of the normalized curve. They are represented by the more abundant taxa in the sample because these taxa have a higher probability of being randomly selected.

We used the mean taxa richness and the taxonomic proportions in the original sample raw counts to subsample the original sample and create the sample's rarified assemblage. This approach reduces the amount of variability that results when the rrarefy function is applied, and produces the assemblage that would occur most frequently with many iterations of rrarefy. A sample's raw counts are first sorted in descending order. The mean richness value generated by the rarefy function is then applied to the descending list of taxa. This identifies the taxa most likely to occur in the rarified sample. Examples of this approach for different raw sample counts are shown in Tables 2-4. In Table 2, the rarefy function indicates a mean richness value of 6 (rounding up from 5.2) for the rarefied sample. The 6 taxa with the highest raw counts become the rarefied assemblage. The proportion (percentage) of each taxon in the raw count is then used to calculate an expected taxon count in the rarified assemblage. It is not possible to count a fraction of an individual taxon (e.g., Table 2 Family A = 33.22 individuals), therefore, each estimated count is rounded up to the nearest whole number in the rarified sample. Rounding up typically results in total counts at or above the prescribed count for the rarified sample (e.g., 100-115 for a prescribed total count of 100). However, richness and diversity metrics for samples that had different raw counts are now comparable.

In some cases, the original sample's taxonomic list cannot be subsampled easily because several taxa near the mean richness cutoff have the same raw count. In Table 3 and Table 4, for example, the count of the last taxon included in the subset assemblage (the count just above the dashed line) has a count identical to that of one or more neighboring taxa. When identical raw counts occur, a random selection without replacement needs to be applied to the affected taxa to avoid bias when determining the mean taxa richness value. Without random sampling of these taxa with identical counts, the program will select the taxon that occurs when the mean richness value is reached (i.e., F in Tables 3 and 4). This creates an unnecessary and odd bias based on a taxon's Latin name - a result of taxa with identical counts being automatically alphabetized. Therefore, a random sample without replacement is applied to taxa with identical counts in order to reach the appropriate number of individuals necessary to meet the richness requirements of the rarefied assemblage. In Table 3, Taxon G was randomly selected to fill the final space in the rarefied assemblage. In Table 4, 3 taxa are needed to meet the prescribed richness value of 6. Random selection picked Taxa E, F, and H. Taxa with the same count values are usually singletons (represented by one individual) or doubletons (represented by two individuals). Therefore, the variability produced by the final, random selection step has minimal influence on most metrics. Richness metrics for specific taxon or taxa may be more influenced by this variability (e.g., EPT richness, Diptera richness, Collector richness).

---

[1] Within the BIBI2 package, the user can specify the taxonomic level (e.g. family, genus, species, etc.) at which the rarefaction function samples the assemblage and can also select the rarefaction sample size (e.g. 100, 200, etc.). We use a standard number of 100 for the Chessie BIBI, and specify the family or genus taxonomic level depending on which metrics are being calculated.

Simplified Explanation

1. The taxa are sorted by raw count in descending order for each sampling event (Table 1).
2. The vegan package function “rarefy” is used to calculate the mean richness for a prescribed sample count (e.g., 100, 200, etc.).
3. The mean richness (rounded up to the nearest whole number) is used as a cutoff for identifying a sample’s taxa with the greatest probability of being randomly selected for the prescribed richness value (e.g., Table 2).
    a. If the count of the final taxon above the threshold is equal to the count of one or more neighboring taxa in the list, a random selection without replacement occurs among these taxa to fill the remaining space(s) in the subset assemblage (Tables 3 and 4).
4. The counts from the new subset assemblage are then represented as a proportion of the new subset’s total, and rounded up to the nearest whole number.
5. Richness and diversity metric are calculated on the rarefied assemblage.

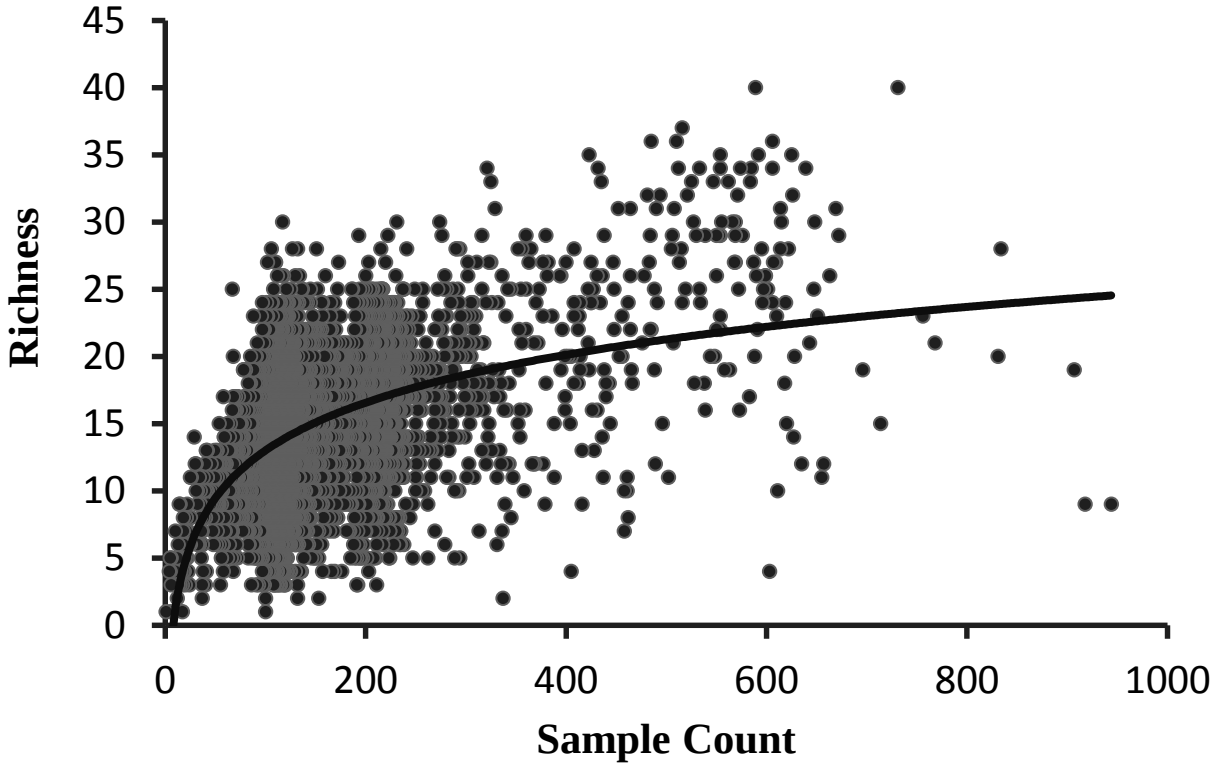

Figure 1. The number of individuals identified (Sample Count) plotted against the number of families observed (Richness). All of the data points (n = 5,271) represent macroinvertebrate assemblages collected in streams with a kick-net in streams of Strahler order ≤ 4 between 1992 and 2009. The trend line depicts the positive relationship between the number of individuals identified and the richness observed.

Table 1. Example of three different rarefied assemblages produced with the function rrarefy from the vegan package for a prescribed total count of 100 organisms. Each rrarefy column represents an iteration of the rrarefy function on the sample's raw count. Richness and diversity metric values are different for each rrarefy iterations. The bottom row is the mean taxa richness found using the function rarefy for the prescribed 100 count subsample.

Taxa	Raw Counts	rrarefy 1	rrarefy 2	rrarefy 3
A	200	38	39	33
B	195	30	30	31
C	190	27	30	36
D	10	3	0	0
E	4	1	1	0
F	3	0	0	0
G	2	1	0	0
H	1	0	0	0
Total	605	100	100	100
Richness	8	6	4	3
Mean Taxa Richness	5.2 (6)			

Table 2. Example of how the rarefy function from the vegan package is used to create the most probable taxonomic assemblage for a rarified sample. In this example, the rarefy function returns a mean richness value of 5.2 unique taxa for a count of 100 individuals. 5.2 is rounded up to 6 taxa and the six most abundant taxa become the taxa comprising the rarified sample ("Subset"). The percentage that each taxa represents in the Subset sample (n = 602) is then used to estimate taxa counts in the rarified sample (Rarified Count). Taxa counts will remain the same each time richness and diversity metrics are calculated on the rarified sample.

Taxa	Raw Count	Subset	Percentage	Rarefied Count
A	200	200	33.22	33
B	195	195	32.39	33
C	190	190	31.56	32
D	10	10	1.66	2
E	4	4	0.66	1
F	3	3	0.50	1
G	2			
H	1			
Total	605	602	100.00	102
Richness	8	6	6	6
Mean Taxa Richness	5.2 (6)			

Table 3. Example of how the most probable taxonomic assemblage for a rarified sample is created when three taxa at the richness cutoff are equally abundant. The rarefy function from the vegan package returns a value of 5.3 unique taxa (rounded up to 6) in a count of 100 organisms, and the dashed line indicates where the assemblage should be subset. Since the program orders the taxa list first by raw count and then alphabetically, Taxa F will always be selected for the rarified sample. This creates an unnecessary bias because Taxa F, G, and H each have a raw count of one and should have an equal chance of being selected. To avoid this bias, the program randomly selects one of the three equally abundant taxa for the rarefied assemblage. In this example, it selected Taxa G.

Taxa	Raw Count	Subset	Percentage	Rarefied Count
A	100	100	36.50	37
B	90	90	32.85	33
C	80	80	29.20	29
D	2	2	0.73	1
E	2	2	0.73	1
F	1			
G	1	1	0.36	1
H	1			
Total	277	275	100.00	102
Richness	5.3 (6)	6	6	6

Table 4. Example of how the most probable taxonomic assemblage for a rarified sample is created when five taxa at the richness cutoff are equally abundant and three need to be selected to create the rarified assemblage. The rarefy function from the vegan package returns a value of 5.3 unique taxa (rounded up to 6) in a count of 100 organisms, and the dashed line indicates where the assemblage should be subset. The 4th through 8th taxa (D – H) each have raw counts of one and should have equal probability of being selected. In this example, the program randomly selects three of the five taxa with raw counts of one to be included in the rarefied assemblage. If three taxa were not randomly selected, the program would favor the taxa that appear first in alphabetical order (i.e., Taxa D F) and create an unnecessary bias.

Taxa	Raw Count	Subset	Percentage	Rarefied Count
A	90	90	37.04	38
B	80	80	32.92	33
C	70	70	28.81	29
D	1			
E	1	1	0.41	1
F	1	1	0.41	1
G	1			
H	1	1	0.41	1
Total	245	243	100.00	103
Richness	5.04 (6)	6	6	6