**Assessing microbial community composition and diversity**

**Problem set 04**

*version March 2, 2025*

**Learning objectives:**

* Gain experience estimating diversity within a hypothetical microbial community

Obtain a collection of “microbial” cells from “seawater”. The cells were concentrated from different depth intervals by a marine microbiologist travelling along the Line-P transect in the northeast subarctic Pacific Ocean off the coast of Vancouver Island, British Columbia.

Sort out and identify different microbial “species” based on shared properties or traits. Once you have defined your binning criteria, separate the cells using the sampling bags provided. These operational taxonomic units (OTUs) will be considered separate “species.”

**Part 1: Description and enumeration**

Construct a table listing each species, its distinguishing characteristics, the name you have given it and the number of occurrences of the species in the collection. Use whatever means to produce a table that allows for easy export in a plain-text file format (such as CSV or TSV) for later import into R for further analysis. The table headers below provide a suggestion for how to format your data collection table.

|  |  |  |  |
| --- | --- | --- | --- |
| Species Number | Species Name | Distinguishing  Characteristics | Number of Occurrences |
| 1 |  |  |  |
| 2 |  |  |  |
| 3 |  |  |  |
| … | … | … | … |

*• Ask yourself if your collection of microbial cells from seawater represents the actual diversity of microorganisms inhabiting waters along the Line-P transect.* *Were the majority of different species sampled or were many missed?*

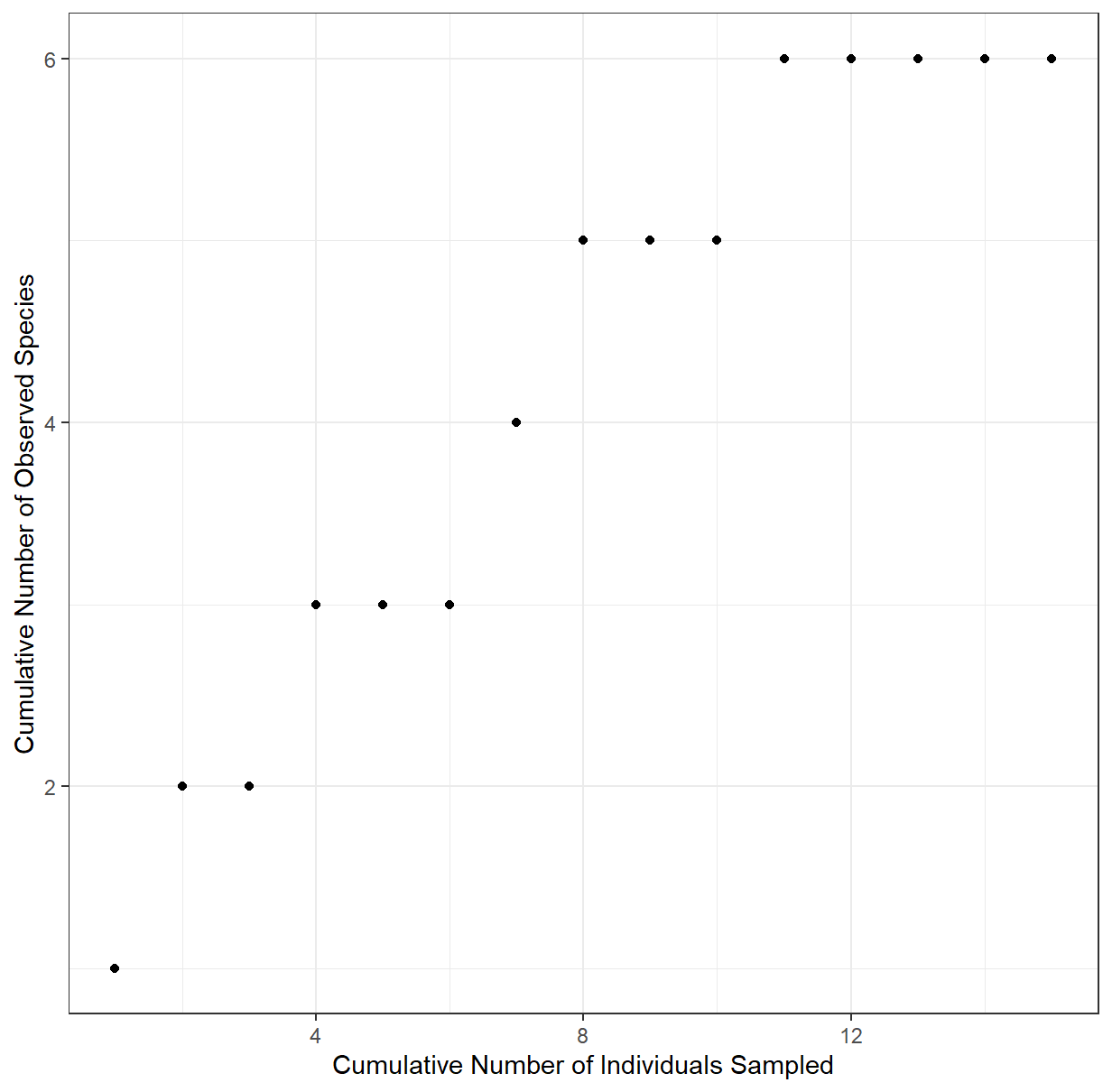
**Part 2: Collector’s curve**

To help answer the questions raised in part 1 you will conduct a simple but informative analysis that is a standard practice in biodiversity surveys. This analysis involves constructing a species accumulation (or collector’s) curve that plot the **cumulative number of species observed** along the y-axis and the **cumulative number of individuals** classified along the x-axis. This curve is an increasing function with a slope that will decrease as more individuals are classified and as fewer species remain to be identified. If sampling stops while the curve is still rapidly increasing then this indicates that sampling is incomplete and many species remain undetected. Alternatively, if the slope of the curve reaches zero (flattens out), sampling is likely more than adequate.

To *manually* construct such a curve for your samples:

1. A cell within your entire collected sample is chosen at random.
2. Record this first cell as your first data point, such that x = 1 and y = 1. This indicates the first data point represents the first cell sampled (+1 on the x axis), and that the species has not yet been sampled (+1 on the y axis).
3. *Without replacement of the first cell*, another cell is chosen at random.
4. Record the next cell as x = 2, since another cell has been sampled (+1 on the x axis). If this next cell is an already sampled species, y remains 1 (+0 on the y axis). Otherwise, a new species is observed, and y increases to 2 (+1 on the y axis).
5. Repeat Steps 3 and 4 until you have proceeded through all collected cells within the sample.

For example…

****

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Species** | **X** | **Y** |
| ***1*** | A | 1 | 1 |
| ***2*** | B | 2 | 2 |
| ***3*** | B | 3 | 2 |
| ***4*** | C | 4 | 3 |
| ***5*** | A | 5 | 3 |
| ***6*** | C | 6 | 3 |
| ***7*** | D | 7 | 4 |
| ***8*** | E | 8 | 5 |
| ***9*** | B | 9 | 5 |
| ***10*** | A | 10 | 5 |
| ***11*** | F | 11 | 6 |
| ***12*** | A | 12 | 6 |
| ***13*** | A | 13 | 6 |
| ***14*** | C | 14 | 6 |
| ***15*** | B | 15 | 6 |
| ***…*** | … | … | … |

**Instead of manually constructing your curve, we can *simulate* randomly sampling from your cell collection from the data you have already collected.**

To get you started, follow the steps below, consulting the documentation as needed to gain practice making use of new functions:

1. Import your summary table generated in Part 1. Ensure your data include species information and their respective frequencies.
2. To ensure each species is represented uniquely, optionally merge the columns containing species name and descriptions using the tidyverse function **unite()**. Then, create a data frame of species repeated according to their occurrence counts using **rep()**.
3. Simulate randomly drawing your cells by randomly rearranging the order of the elements using the function **sample()** to generate a vector of drawn cells.
4. Make a new vector that represents the number of unique species observed as you iterate through the object created in Step 3. This is a bit more challenging; thus, we provide the code, where **sample\_vector** represents the vector created in Step 3:

**observed\_species <- numeric(length(sample\_vector))**

**for (i in 1:length(sample\_vector)) {**

**observed\_species[i] <- length(unique(sample\_vector[1:i]))**

**}**

1. Create a data frame for plotting bringing together the cumulative individuals sampled (the X coordinates—a vector of consecutive numbers) and the vector from Step 4 (the Y coordinates). See the example data and plot above.
2. Plot the collector’s curve using **ggplot()**with **geom\_point()**.

• *Paste in the R code and your generated plot below to be submitted as part of your assignment.*

• *Does the curve flatten out? If so, after how many individual cells have been collected?*

• *What can you conclude from the shape of your collector’s curve as to your depth of sampling?*

**Part 3: Diversity estimates**

Using the table from Part 1 calculate species diversity using the following indices or metrics:

Simpson Reciprocal Index

fractional abundance of the species (

For example, for a sample of two species with five individuals each,

The higher the value is, the greater the diversity. The maximum value is the number of species in the sample, which occurs when all species contain an equal number of individuals. Because the index reflects the number of species present “richness” and the relative proportions of each species with a community it provides information on the “evenness” of the population structure. Consider that a community can have the same number of species (equal richness) but manifest a skewed distribution in the proportion of each species (unequal evenness).

• *What is the Simpson Reciprocal Index for your sample? Instead of manually coding this calculation, we will make use of the R package* ***vegan****.**This R package**contains a large collection of functions useful for analyzing microbial community data. To calculate the Simpson Reciprocal Index of your sample, use the function* **diversity()** *from this package. Show your code and the resulting value below.*

Another way to calculate diversity is to estimate the number of species that are present in a sample based on the empirical data to give an upper boundary of the richness of a sample. Here we use the Chao1 richness estimator.

total number of species observed

number of species observed once

species observed twice or more

• *What is the Chao1 estimate for your sample? Use the function* **estimate()** *from the* **vegan** *package. Show your code and resulting value below.*

**Part 4: Comparing samples**

While diversity is one thing, distinctiveness is another matter. Thus, another important perspective in ranking sites is how different microbial communities are from one another. One of the simplest metrics to measure community similarity is the Jaccard coefficient of community similarity, to contrast distinctiveness between all possible pairs of sites:

the number of species common to both communities

the total number of species present in the two communities

For example, if one sample contains only 2 species and the other sample contains 2 species one of which is common between samples, the total number of species present is 3 and number shared is 1, so 1/3 = 33%. This index ranges from 0 (no species in common) to 1 (all species in common).

• B*uild a* CC*j matrix based on pair-wise similarity for each sample (Sample 1 vs 2, Sample 1 vs 3, Sample 1 vs 4, etc) using the function* **vegdist()** *from the* **vegan** *package. The input needed is a community data matrix where each row represents a sample, each column represents a species, and the values represent the occurrences (abundance) of the species in each sample. Show your code and complete the* CCj *matrix below.*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Sample 1 |  |  |  |  |  |
| Sample 1 | 1 | Sample 2 |  |  |  |  |
| Sample 2 |  | 1 | Sample 3 |  |  |  |
| Sample 3 |  |  | 1 | Sample 4 |  |  |
| Sample 4 |  |  |  | 1 | Sample 5 |  |
| Sample 5 |  |  |  |  | 1 | Sample 6 |
| Sample 6 |  |  |  |  |  | 1 |

*• Based on the matrix above, what does this tell you about the six samples analyzed by the class?*

**Part 5: Concluding activity**

• *How does the measure of diversity depend on the definition of species in your samples?*

• *Can you think of alternative ways to cluster or bin your data that might change the observed number of species?*

• *How might different sequencing technologies influence observed diversity in a sample?*