International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor 2024: 2.102

From Data to Insights - Biodiversity Analysis for Emerging Researchers

Dr. Ashvini Kumar Joshi[1], Dr. Sanjay Tomar[2], Dr. Aditya Sharma[3]

[1]Wildlife and Conservation Laboratory, Department of Zoology, M.L.V. Government College, Bhilwara (Raj) 311001
Corresponding Author Email: *kashvini80[at]yahoo.com*

[2]Departmemnt of Botany, S.P.C. Government College, Ajmer (Raj.) 305001
Email: *sanjaytomar.ajmer[at]gmail.com*

[3]Departmemnt of Botany, S.P.C. Government College, Ajmer (Raj.) 305001
Email: *adityasharma2876[at]gmail.com*

Abstract: ***In the early stages of biodiversity research, many researchers face challenges in identifying the appropriate statistical analyses required to obtain meaningful results and guide their studies effectively. Often, data is collected over an extended period, only for researchers to realize during the analysis phase that certain parameters or aspects were overlooked due to a lack of awareness about post-field data analysis techniques. This article aims to assist researchers in the collection and analysis of biodiversity field data. While experienced researchers may explore a vast array of analytical possibilities, this article will peculiary useful for guiding less experienced researchers in the foundational aspects of biodiversity data analysis and to develop an insight to see across the data.***

Keywords: statistical analyses, early researchers, biodiversity data

1. Introduction

It usually becomes very much confusing for a researcher of biodiversity to statistically analyze field data. The data without well analysis are not of use as the result always remains incomplete. Understanding field data is very crucial for conservation, ecosystem management, and ecological research. It develops an understanding about the biodiversity patterns of Species Richness and Abundance across ecosystems and provides a deep insights into the relative proportions of species and their roles in the ecosystem. It monitors change in trends by comparing data across time such as species decline or introduction of invasive species and identifies biodiversity hotspots and areas of low diversity. A proper analysis assess and evaluates the impact of urbanization, deforestation, agriculture, and other human-induced changes on biodiversity and also assesses how shifting climatic conditions affect species distribution, abundance, and interactions. Moreover it helps pinpoint vulnerable species and ecosystems at risk of extinction or degradation with measurement of the success of conservation efforts, such as habitat restoration or protected area designation. Other importance of statistical analysis is hypothesis Testing, Predictive, Policy and Decision-Making as Evidence-Based Policies, Economic Valuation, Interdisciplinary Insights and Big Data Analysis. The article will provide an initial guideline to the student to understand the type of statistics; they can apply to analyze their data. The article has been divided in three section General statistics, Ordination analysis and Biodiversity indices containing general aspects of data analysis.

2. General Statistics

1) **Measures of Central tendency**- A researcher can calculate Mean (Arithmatic) and Median to know the central tendency of its data. These can be calculated for richness/abundance data for a given species or group of species across sites or sampling units. One can use mean when the analysis requires a measure of central tendency or when the score in a distribution are likely to be grouped around a central point. When the extreme observations distort the mean one should use median because the mean reflects extreme values but the median does not.
2) **Measures of variability**- These measures suggest the dispersion of individual observations around the mean of a large series of observations. When the observations are equal, the variability will be zero and when the observations are unequal, it becomes positive. Some measures are-

2.1 Mean Deviation

It is an average mean of deviation of values from central value. If the deviation is greater than the mean, the deviation is positive and if less than the deviation is negative. The mean deviation can be calculated about any of the three averages i.e. mean, median and mode.

2.2 Standard Deviation

It is most widely used and the best measure of dispersions among all other measures as it is less affected by the fluctuations of sampling. A large standard deviation show that the measurement of frequency distribution are widely spread out from mean while small standard deviation shows close spread in neighbourhood of mean. It is the square root of variance and helps to find the standard error. A population's variability is gauged by the standard deviation. It shows the range of varieties and indicates the distance of the individual variety from the mean of population (Annadurai, 2015).

2.3 Variance

This is a standard measure of variation and directly related to the standard deviation. The variance is the arithmetic mean of the square of sums of the deviations from the mean value of the data and it is described as the square of Standard Deviation. It indicates the variability clearly and it is most informative among the measure of dispersion for populations.

3. Interferential statistics/ Test of Significance

Data analysis and decision-making and comparison are done through the application of statistical tests. These can be broadly classified as parametric and non-parametric tests. Paramatric test generally requires larger sample, data with normal distribution, based on intervals and ratio scales of measurement (eg. height, temperature etc.). t-Test, F-Test, Chi-square test are examples of such tests. The non-paramatric tests may be used with small samples, data based on nominal and ordinal scale of measurement (rank, rating etc.). The examples of such tests are Mann-Whitney U Test, Kruskal-Wallis Test etc. As normal distribution of plants and animals is not confirmed by any /distribution/abundance data, non-paramatric tests are appropriate for ecological studies (Henderson, 2008).These tests are being described in brief below-

3.1 T- test

It compares mean of two groups if they are significantly different. P value indicates the significant difference. If it is less than 0.05 than it is statistically significant and the null hypothesis should be rejected. P-value greater than 0.05 indicates result is not statistically significant and the null hypothesis is not rejected. A smaller p-value is more significant than higher p-value. The probability of alternative hypothesis becoming true is (1-P value).The test is preferred normally when the sample size less than 30.

3.2 F- Test

This is a significance test used to know if difference is found between the variance of two population and samples.

3.3 ANOVA (Analysis of Variance)

Used to compare a species' mean abundance among three or more groups (such as treatment groups or different places).

One-way ANOVA is used when comparing the mean abundance of groups based on a single factor, for example bird abundance across various vegetation types. Two-way ANOVA: It is used to examine interaction of two factors, such as the impact of habitat type and season on bird abundance.

3.4 Mann-Whitney U Test

When the median of two groups/sites are significantly different one can use this test. It assumes that the samples are drawn independently and randomly from their respective population and the test population do not differ except their median.

3.5 Kruskal-Wallis Test

It is s non paramatric test used to compare three or more samples. It is a non-pa- ramatric alternative of one-way ANOVA (Rastogi, 2017) It is used when the data for analysis consists only of ranks and analogous to the F-test used in analysis of variance.

3.6 Correlation

It is used to measure the strength and direction of relationship between two quantitative variables (Baldi and Moore, 2012). The movement in one variable is accompanied by corresponding change in other variable also. It may be positive and negative and helpful to study direction and nature between two or more variables (Rastogi, 2017).It is measured by scatter plot and Correlation coefficient by Karl Pearson's method and Correlation coefficient by Spearman's rank method.

3.7 Regression Analysis

It is a statistical analysis used to ascertain how strongly a dependent and independent variable are related to one another in a particular population. The relationship is expressed as a curve called regression curve. It is used to predict value of one variable from the value of other variable and the value of correlation coefficient suggests the degree of association of two variables but regression analysis says how change in one variable can affect change in other variable. If a researcher is interested in understanding how environmental factors (such as temperature, habitat structure or vegetation cover) influence suppose bird abundance, it can use regression models.

3.8 Ordination analysis

Ordination analysis is a set of statistical techniques used to explore and visualize the relationships between multiple variables (such as species or samples) in ecological or environmental data. Ordination methods help to minimize the dimensionality of complex datasets, allowing researchers to uncover patterns in the data and interpret the structure of biodiversity or ecological communities. These are particularly useful in ecology for visualizing and analyzing species composition, environmental gradients, and ecological relationship, identifying patterns and trends for investigating how different factors (e.g., environmental variables, treatments, or species) relate to one another across multiple sites or time points? We will discuss commonly used ordination analysis methods in this section.

1) Principal Component Analysis (PCA) -It is an unsupervised linear ordination approach. By determining which primary axes (or components) best captures the range of the data, it reduces the dimensionality of a dataset without losing actual information (Dar, 2021). It finds the principal

components,that account for the majority of the dataset's variance. The most significant variance is typically captured by the first few primary components, enabling a simplified 2D or 3D presentation. It works well with continuous data, such as abundance data. It can be applied to both environmental and species data and visualize the general structure and trends in community composition, especially when there are a number of factors. In order to determine how the places cluster according to their species composition, It can be used to examine species abundance across various locations. The first principal component (PC1) explains the largest variance, and the second principal component (PC2) explains the next largest variance. Points on the plot are samples or species, and their positions relative to each other reveal patterns and gradients.

2) Non-metric Multidimensional Scaling (NMDS)- It is an unsupervised, non-linear ordination method. The objective of NMDS is to represent complex, multi-dimensional data in fewer dimensions, maintaining the rank order of dissimilarities or distances between samples. NMDS has flexibility and can select one of the many ordination technique uses a dissimilarity matrix such as Bray-Curtis for abundance data and Jaccard's for biography data to lessen the dimensionality of the data*. It places samples in a low-dimensional space while minimizing the stress (i.e., the difference between the distance matrix and the distances between points in the ordination).It is used with community data (species abundance or presence/absence) and ecological distance measures and suitable for non-linear relationships between variables. NMDS is commonly used to visualize species composition data across different ecological sites to explore community similarity or dissimilarity. In NMDS, the distance between points in the ordination plot reflects the dissimilarity between samples. Samples and sites that are close together have similar species compositions, while distant points indicate more dissimilar communities. Stress Value is a measure in NMDS. If the stress values is low, it indicates better fit to the original distance matrix. A value below 0.2 generally considered good.
3) Correspondence Analysis (CA)- It is a common Principal Component Analysis designed to analyze qualitative data (Abdi et.al., 2010) available in contingen- cy tables (species × sites) or frequency data (e.g., species counts) to examine the relationship between rows (e.g., sites) and columns (e.g., species). It is particularly useful for species abundance data or when examining species-environment relationships. It produces a map of the table in which there is a point for each row and each column (Kroonenberg et.al., 2004). It can be used to analyze to analyze variation in species composition varies across different habitats or environmental conditions, such as soil pH, moisture, or temperature. It focuses on maximizing the relationship between samples and species. The axes represent gradients of species abundances, and the position of samples or species in the ordination space indicating their relationship to these gradients.
4) Canonical Correspondence Analysis (CCA)- It is a supervised ordination method. It is a multivariate method that combines ordination with multiple regression, allowing researchers to model the relationship between species composition and environmental variables by assuming a common response model to all species (Braak, 1986). CCA uses a linear relationship between species data and environmental variables to create an ordination, revealing how species compositions are influenced by environmental factors. It is suitable for species-environment data, specifically when one wants to explicitly relate environmental gradients to species composition. The ordination plot shows how species are influenced by environmental gradients, with species and sites plotted in relation to the environmental variables.
5) Principal Coordinates Analysis (PCoA)- It is unsupervised, linear ordination method. It is similar to PCA but used to know non-eucledean like Bray-Curtis dissimilarity which is commonly used to describe pairwise dissimilarity between samples**.It is commonly used in genomics, ecology and microbiology. PCoA reduces the dimensionality of a distance matrix while retaining the relationships between samples based on their dissimilarities. Thus it can be used to analyze genetic distance data or ecological dissimilarity between different habitats or species communities. The PCoA represents as a scatter plot where points closer to each other show more similarity in species composition or community structure.

Ordination methods are powerful tools in community ecology, enabling researchers to explore and interpret complex ecological patterns. Which ordination method would be used depends on the research question, the type of data (e.g. species abundance, presence/absence, dissimilarity) and the underlying ecological processes.

Diversity Indices

One of the main challenges in studying biodiversity is quantifying and comparing the diversity of biological communities across different spatial and temporal scales. This is where biodiversity indices come into play. Biodiversity indices are mathematical tools that allow researchers to measure and quantify the diversity of species within a community or ecosystem. These indices help provide a numerical representation of biodiversity, which can be used to compare different habitats, monitor changes in biodiversity over time, or assess the impact of environmental factors and human activities. By translating complex ecological data into a single value, biodiversity indices make it easier to interpret patterns of diversity and identify areas of concern, such as biodiversity loss or habitat degradation. There are different types of biodiversity indices, each focusing on different aspects of biodiversity, such as species richness, evenness, or the distribution of species. The most commonly used indices can be broadly categorized into alpha diversity indices, which measure diversity within a single community or site, and beta diversity indices, which measure differences in diversity between multiple sites or communities. Additionally, gamma diversity refers to the overall diversity within a landscape or region, encompassing both

Paper ID: SR25112094537 DOI: https://dx.doi.org/10.21275/SR25112094537

alpha and beta diversity. In this below section indices of biodiversity are being discussed

1) Species Richness Indices: It is the total number of species present in a given area. A simple count of species, but it does not account for the abundance of each species or how evenly they are distributed. The common indices are-
a) Margalef's Diversity Index (D_{Mg}) and Menhinick's Index (D_{Mn}): These are simple species richness indices which can be calculated easily (Clifford and Stephenson, 1975; Whittaker, 1977). These are meaningful indices useful in biodiversity estimation.

$$D_{Mg}=\frac{(S-1)}{\ln N}$$

$$D_{Mn}=\frac{(S)}{\sqrt{N}}$$

Here S = Number of species recorded and N=Total no. of individuals in the sample

The Higher values indicate greater species richness relative to the number of individuals, while lower values suggest lower species richness or more individuals per species. The Margalef index is calculated as S-1 and the Menhinick's with S (Magurran, 2013).

2) Alpha Diversity Indices-Alpha diversity refers to the diversity within a particular area or ecosystem, and it is typically measured by various biodiversity indices. These indices aim to quantify the variety of species in a given location, accounting for both the species number (species richness) and their relative abundance (evenness). The principal biodiversity indices used to assess alpha diversity:

a) Shannon-Wiener Index (H')
A regularly used index that observes both species richness and evenness. The value of the index is 0 to infinity. A higher value specifies greater diversity with more even distribution of individuals across species. ,

$$H' = -\sum p_i \ln p_i$$

Where *pi* is the number of individuals of each species (n) divided by the total number of species in the sample (N) = n/N

b) Brillouin Index (HB)
It is used when the randomness of a sample is not confirmed or the community is completely censused and every individual accounted for. (Pielou, 1969,1975).Both Shanon and Brillouin index give same estimate of diversity but Brillouin index give lower value as it measures both sampled and unsampled portion of the community (Magurran, 2013).

$$HB = \frac{\ln(N!) - \sum \ln(n_i)!}{N}$$

Where N is the total number of individuals, n_i is the number of individuals of the i_{th} species#

c) Simpson's Index (D)
It Measures the probability that two randomly selected individuals from a large community will belong to the same species. It focuses more on the most abundant species in the sample and less sensitive to species richness (Magurran, 2013) common and dominant species. Its values range from 0 (infinite diversity) to 1 (no diversity). The higher value of D denotes lower diversity so it is usually expressed as 1-D. a higher value would indicate greater diversity. it is more meaningful but, less popular, indice of biodiversity measurement.

$$D=\sum\left(\frac{n_i\,(n_i-1)}{N(N-1)}\right)$$

Here n_i is the number of individuals in the i_{th} species and N is the total number of individuals (Magurran, 2013)

d) The Berger-Parker Index
It is a simple dominance measure (Berger and Parker, 1970) that is easy to calculate. It expresses proportional abundance of the most abundant species. It can be used in the community dominated by many species (kim)

$$d= N_{max}/N$$

Where N_{max} is the number of individuals in the most abundant species

e) Evenness Indices (J')
The evenness index measures the even distribution of individuals across the species in a community. The value of evenness index ranges from 0 (uneven distribution) to 1 (perfect evenness i.e. all species are similarly abundant). There are many evenness indices Shanon evenness measure, Carmargo's evenness index, Smith and Wilson's evenness index, Simpson evenness measure etc. (Magurran, 2013) concludes Smith and Wilson's evenness index as most satisfactory index and recommended Simpson evenness measure also.

f) Chao1 Index
A non-parametric estimator of species richness, particularly useful when some species are rare or not observed in the sample. It is calculated based on the number of singletons, species found only once and doubletons, species found twice (Magurran, 2013). It provides an estimate of the total species richness, including those that were not observed in the sample.

Each index gives a slightly different perspective on the diversity of a community, and they can be used in combination to get a more comprehensive understanding of alpha diversity.

4. Beta Diversity Indices

Beta diversity refers to the variation in species composition between different ecosystems, habitats, or communities. It measures the turnover or differences in species diversity between sites, indicating how much species composition changes from one location to another. Beta diversity indices help to quantify these differences and are essential for understanding spatial patterns of biodiversity. Here are the principal biodiversity indices for beta diversity:-

4.1 Bray-Curtis Dissimilarity Index

One of the most commonly used indices to measure beta diversity. It is commonly known as a modified Sorensen's index and calculates dissimilarity between two sites based on species abundance (Hao et.al. 2019). The ranges of the index is from 0 (sites are identical) to 1 (completely dissimilar). A higher value indicates greater dissimilarity between the two communities. The Bray-Curtis index is often used in ecological studies involving species abundance or relative abundance. Clarke and Warwick (2001a) tested the index using six standard criteria and found that the Bray-Curtis index meets all criteria well. The formula to calculate the index is-

$$C_N = \frac{2jN}{(N_a+N_b)}$$

Here, N_a= Total number of individuals in site A
N_b= Total number of individual in site B
2jN=the sum of the lower of the two abundances for species found in both species

4.2 Jaccard's index

It is a measures the similarity between two sets/sites based on species presence/absence, without considering species abundance and commonly used in community ecology to compare species composition between sites. It ranges from 0 (no species in common) to 1 (same species in both sites). Higher values indicate greater similarity in species composition.

$$Cj = \frac{a}{a+b+c}$$

Here a= is the total number of species present in both samples, b is the number of species present only in sample 1 and c is the number of species present only in sample 2 (Magurran, 2013).

4.3 Sorensen Index

Measures the similarity between two sites based on species presence or absence, similar to the Jaccard Index, but gives more weight to shared species between sites, emphasizing species that are common to both locations. It's value also ranges from 0 (no shared species) to 1 (identical species composition). A higher value indicates greater similarity between sites. It is calculated by twice the number of shared species divided by the sum of species in both locations/sites (##).

$$Cs = \frac{2a}{2a+b+c}$$

4.4 UniFrac Index (Phylogenetic Dissimilarity)

A phylogenetic measure of beta diversity detecting evolutionary relationships between species. It is commonly used to characterize microbial community and to determine if the microbial communities are different significantly or not. Sets of taxa in a phylogenetic tree are measured to reveal the phylogenetic distance (Lozupone and Knight, 005).Values range from 0 (no phylogenetic difference) to 1 (completely dissimilar phylogeny). Higher values indicate greater phylogenetic dissimilarity.

5. Conclusion

Each of these technique offers different insights into the composition and turnover of species across sites, and the choice of index depends on the data e.g., presence/absence vs. abundance and the research goals e.g., phylogenetic vs. ecological dissimilarity. Nowadays, numerous applications and software are available that can perform data analysis quickly and with high accuracy. Examples include SPSS, EstimateS, MATLAB, Biodiversity Pro, and R programming, which are also effective for data visualization. The selection of a statistical technique depends on the researcher's preference, but having knowledge of the appropriate method can save time and effort while delivering optimal results that benefit both the researcher and society.

References

[1] B. Annadurai, "A textbook of Biostatistics. New Age International (P) Limited., Publisher, New Delhi pp 1-368

[2] Baldi, B. and Moore, D. S. (2012). The Practice of satistics in the Llife science Published by W. H. Freeman and Company, New York pp1-712.

[3] Berger,W. H. and Parker,F. L. (1970) Diversity of planktonic Foraminifera in deep sea sediments. Science 168,1345-1347.

[4] Braak, Cajo J. F. Ter. (1986). Canonical Correspondence Analysis: A New Eigenvector Technique for Multivariate Direct Gradient Analysis. *Ecology*, *67*(5), 1167–1179.

[5] Clifford,H. T. & Stephenson, W. (1975). An introduction to numerical classification. London: Academic Press.

[6] Emden,H. F. V. (2008). Statistics for Terrified Biologists. Blackwell Publishing. Pp 1-338.

[7] Fowler, J. and Cohen,L. (1987). Statistics for Ornoithologists. British Trust for Ornithology Guide 22.

[8] Henderson, P. A. (Practical Methods in Ecology Blackwell Publishing Maldane, USA. Pp 1-163.

[9] Hao, M., Corral-Rivas, J. J., González-Elizondo, M. S. *et al.* Assessing biological dissimilarities between five forest communities. *For. Ecosyst.* **6**, 30 (2019).

[10] Kroonenberg, Pieter & Greenacre, Michael. (2004). Correspondence Analysis. 10. 1002/0471667196. ess6018.

[11] Lozupone C, Knight R. UniFrac: a new phylogenetic method for comparing microbial communities. Appl Environ Microbiol 2005; 71: 8228–35. 10. 1128/AEM. 71. 12. 8228-8235. 2005

[12] Magurran, A. E. (2013). Measuring Biological Diversity. John Wiley and Sons, New Jersey.

[13] Pielou, E. C. (1969) An introduction to mathematical ecology. NewYork:Wiley

[14] Pielou,E. C. (1975) Ecological diversity, NewYork: Wiley InterScience.

[15] Rastogi, V. R. 92017). Biostatistics. Published by Scientific International Private Limited, New Delhi. Pp 1-471.

[16] Whittaker, R. H. (1977) Evolution of species diversity in land communities. Evolutionary Biology 10,1-67.

[17] https://faculty. tnstate. edu/tobedeleted_2014_0227/ ganter/B412%20L16%20Communities. html

[18] https://decodingbiosphere. com/ecology2/understanding-beta-diversity/

[19] https://library. virginia. edu/data/articles/starting-non-metric-multidimensional-scaling-nmds

[20] https://www. geeksforgeeks. org/principal-coordinates- analysis-pcoa-a-comprehensive-guide/