1/24/2024

* Data screening
  + Examine summary statistics – look for unrealistic values
  + Missing data
    - Lost data
    - Unable to take a measurement (broken equipment, no response, poor field conditions)
* Dealing with missing data
  + **Complete case**: omit any case that is missing data
    - Discard information
    - Decrease sample size
    - Non-random removals
  + **Imputation**: “filling in” missing data with plausible values
    - Regression models, Monte Carlo simulations
    - Still not “real data” and can bias estimates
      * Need to include in discussion of paper or report
    - Interpolation between nearest values
* Data transformations and standardizations
  + Multivariate normality and model assumptions
    - Assumption of normality for many univariate stats
    - Same for many multivariate statistics
    - Thinking of normality, but in a 3-dimensional space
  + Many tests, rarely used (e.g., Mardia’s, Henze-Zirkler’s, Royston’s)
    - Test each variable for univariate normality (in 2 dimensions)
    - If any variable is not normally distributed, no multivariate normality
    - But the opposite is not true (multivariate normality does not mean that all the variables are univariately normal)
  + Data transformation
    - Y\* = f(Y): Y\* = transformed variable, f = mathematical function, Y = original variable
    - Improve assumptions of **normality, linearity, homogeneity of variances**
    - Patterns of transformed data may be easier to understand and communicate than patterns in the raw data (e.g., species-area relationship on log-log axes, qPCR efficiency on log-DNA copy number (x) by Cq value (y))
    - Log transformation:
      * X must be ≥ 0 (need to add a 1 to zero values)
      * Compresses high values and spreads low values
      * Equalizes variances when mean and variance are correlated
      * When ratio of largest to smallest is >10
      * Highly positively skewed data (bulk of data is in lower values of x)
    - Square root transformation:
      * X ≥ 0
      * Similar, but less dramatic than log trans
      * Used with count data
      * Count data follow Poisson distribution (i.e., mean = variance)
      * Yields variance independent of mean
    - Arcsine square root transformation:
      * Spreads out values in middle
      * Good for data bounded between 0 and 1
      * **Inappropriate for binomial data or probabilities**
      * **Should be proportion or percent data**
    - **Transformations summary**
      * Use **log** or **square root** for highly skewed data or data ranging > 2 orders of magnitude
      * **Arcsine square root** for proportions
      * Transformations applied to a related variable set should use the same transformation
      * Unrelated variables can be transformed independently
  + Data standardizations
    - Facilitate comparison and analysis of variables with unequal and variable sample units
    - Applied to rows or columns of data matrix
    - Column standardization
      * Adjust for differences among variables
      * Focus is on the profile across a sample unit
      * E.g., different environmental variables with different units
    - Row standardization
      * Adjust for differences among sample units
      * Focus is on the profile within a sample unit (e.g., get relative abundance of species within a site)
    - Column Z-score standardization:
      * Converts data to z-scores (mean = 0, variance =1)
      * Places variables on equal footing
      * Necessary when variables have different scales or units of measurement
      * Putting data into units of standard deviation
    - Column total standardization:
      * X >= 0
      * Commonly used to adjust for unequal abundances among species
      * Relative abundance across sample units/sites
    - Row total standardization:
      * Commonly used to adjust for unequal abundances among sites
    - When to standardize?
      * Depends on variability among rows and columns as measured by the coefficient of variation (cv)
      * Cv = sd/mean
      * Table 9.2 (McCune and Grace 2002)

|  |  |
| --- | --- |
| CV (%) | Variability among rows or column |
| <50 | Small. Relativization usually has small effect |
| 50-100 |  |
| 100-300 |  |
| >300 | Very large |

* Outliers
  + Outliers are recorded values of measurements or observations that are outside the range of the bulk of the data
  + Outliers can have dramatic effects on statistical tests by increasing the variance in the data
  + Univariate outliers
    - Examine standard deviation scores (z-scores) for each variable separately
  + Multivariate outliers
    - Examine deviations (z-scores) of the sample average distances to other samples (e.g., Bray-Curtis distance)
    - Extreme observations are those >3 standard deviations
  + Rules for checking outliers
    - Examine data at all stages (e.g., raw, transformed, standardized, distance)
    - Know impacts of outliers for given analysis
    - **Delete extreme values only if it makes sense**
    - Conduct sensitivity analysis to quantify the influence of outliers (calculate with and without outliers and see what the influence is)
* Lectures posted on Google Drive

1/26/24

* Distance vs dissimilarity
  + P-dimensional data
  + Collection of points forms a data cloud
  + The shape, clumping, and dispersion of the data cloud contains information we seek to describe
  + How similar (or dissimilar) are the different variables
  + **Similarity** = the ratio of the number of attributes two objects have in in common to the total list of attributes between them
    - Objects that have everything in common have a similarity of 1
  + **Dissimilarity** = the number of attributes two objects have uniquely
    - 1-similartiy
  + **Distance** is a geometric conception of the proximity of objects in high-dimensional space defined by measurements on the attributes
  + Distances and dissimilarities are often used interchangeably
    - Dissimilarities are bound between [0,1]
    - Distances are usually unbounded on the upper end
* The metric used will be dictated by the type of data, question, or analysis
  + Choice is guided by personal preference when several metrics are plausible
  + Always good to test out several metrics to see how they may influence interpretation
  + With new kinds of data, do a literature review to see what metrics others have used
* **Euclidean distance**
  + Most commonly used distance metric
  + Data are usually column standardized
  + Used in PCA
  + Basically the Pythagorean theorem
  + Emphasizes outliers
  + **Not appropriate for presence/absence data or count data with many zeroes**
* **Manhattan (city-block) distance**
  + Compared to ED, gives less weight to outliers (no squared differences)
* Proportional (Dis)similarity metrics
  + Express (dis)similarity as a ratio of shared (C) to total (A,B,C)
  + Sorensen’s Index
  + Bray-Curtis
    - Bray-Curtis is for continuous data but collapses down to Sorensen’s with ones and zeroes
  + Maximum when no species are shared (0 = identical)