Practical Stats Newsletter for October 2020

Subscribe and unsubscribe: http://practicalstats.com/news
Archive of past newsletters
http://practicalstats.com/news/archive.html

In this newsletter:
A. Practical Stats Courses
B. Multivariate Graphs 1: Principal Components Analysis
C. Free Resources

A. Practical Stats Courses
On our online training site: https://practicalstats.teachable.com/

Our Nondetects And Data Analysis (NADA) course is available online. It's a complete coverage of data analysis with nondetects and 'remarked data': summary statistics, regression, group testing, trend analysis and even some multivariate methods, all without substituting fabricated numbers like ½ the detection limit. One year's access to the materials costs $795. The R scripts included provide 37 new functions to make data analysis easier, and are a step forward from the NADA package in R.

Our Applied Environmental Statistics courses cover methods from simple statistics through trend analysis. They are also an introduction to using R software, the most widely used statistics software in the world. They are available in two parts, each $650 USD for a 1-year access for one person. Or get both courses together in a bundle for $1200 USD. See our online training site at the link above.

B. Multivariate Graphs 1: Principal Components Analysis
How can one make sense of what's going on in a multivariate maze of observations? What patterns and structures are evident? Are there groupings of data that are important? These questions can be answered by looking at two-dimensional graphs that present information available in data of many more dimensions.

Principal Components Analysis (PCA) is a mathematical projection of multivariate data onto two or more dimensions. PCA defines linear axes through the maze of observations that follow the directions of maximum variation. The first principal component describes the "longest direction" of the data. If the cloud of data were cigar-shaped, the first principal component would be the axis down the center of the cigar's length. The second principal component is computed in the same way given the constraint that the second axis must be perpendicular radially around the first axis -- orient the data so the first component is horizontal and spin the data around the first axis to find the view that maximizes the vertical variation of data. Project the data onto the plane described by the first and second components and you have a view that maximizes data variation as a slice through multidimensional space.

The two-dimensional PCA plot can be thought of as a plane slicing through the center of the data cloud. On the plot are projections of data locations onto the two-dimensional slice. What the two-dimensional plot does not pick up are the actual locations of data in the other dimensions {for example, the distances in front of or behind the figure's flat surface). The plot does not show all of the information in the k dimensions of the data and so loses information in the process. But it is the "best" 2-D slice.

PCA is not a hypothesis test -- "its just math". Like any plot it provides the investigator with ideas of what the data look like, and a way to show important groupings or patterns in the data to others. It may suggest important influences on the data, but it doesn't prove them. Plotting axes of each of the k original variables along with data locations is called a biplot and is the most common plot produced using PCA. All variable axes intersect at the (0, 0) center, where the center represents the mean of each variable. Figure 1 is a PCA biplot of 19 sampling locations in an estuary (Warwick, 1971) shown as numbers, along with vectors representing the k=6 physical/chemical characteristics of the locations used to compute the PCA. The horizontal and vertical axes are the 1st and 2nd principal components.

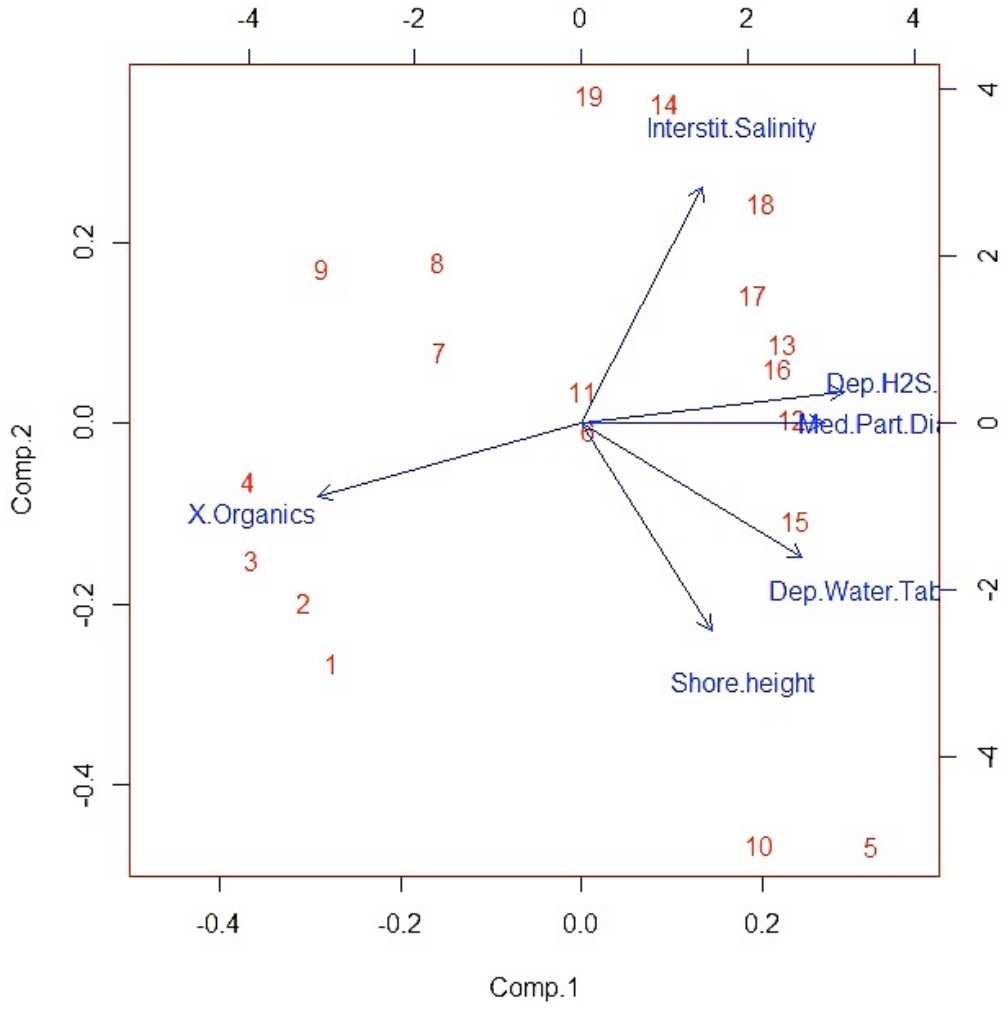

Figure 1. PCA biplot of physical/chemical characteristics at 19 sampling locations.

Variables whose arrow heads point toward data locations have high values for those locations. Sites 3 and 4 have high %organics content (X.Organics) in their sediments, for example. Variables whose vectors point nearly 180^o away from the location have low values for that location. Sites 13 and 16 have low %organics and high median particle diameters (Med.Part.Di) of their sediments. Sites at the top of the plot have higher interstitial salinities than other locations. The first (horizontal) component appears to describe a gradient with organic and more clay rich sediments to the left while to the right are sandy sediments with larger depths to the sulfide layer (Dep H2S) and so are more oxic.

There are four equally-valid arrangements of the algebraic signs of the X and Y scales on a PCA biplot. You could be looking at the 2-D plane of the plot from the back or the front so data on the right (+) side in one plot would be on the left (–) side from the opposite vantage point. In addition, the data in the top half (+) could be on the bottom half (–) if you were standing on your ceiling with your head towards the ground. You as the viewer could stand in four different vantage points in multivariate space and look perpendicularly at the same plane. Different software may choose a different vantage point and so plot observations switched in algebraic signs. All four plots show the same information and are interchangeable. In addition, the scales along the components are arbitrary. Some software could return "scores" - the values along the first and second components for each observation - that are ten times or one-third, etc. of those of other software. The plots will be identical but the numbers along the component axes differ.

PCA is not a parametric method, there is no assumption of normality for any variable, but it is a linear and metric (uses the scales of input variables) method. This produces two considerations:
1. Transformations are sometimes taken if a variable's data are strongly skewed as the linear axes will otherwise shoot out to a large outlier and ignore variation within the bulk of the data. Taking logs or cube or square roots of a variable is standard in order to down-weight the effect of its outliers.
2. If one variable has higher numbers (500 to 10,000) while a second variable such as a concentration has lower numbers (0.01 to 1.9) the one with higher numbers will dominate the effect of the one with lower numbers. To give all variables the same potential influence they are first standardized so all have a mean of 0 and standard deviation of 1 before computing PCA. This is called 'using the correlation matrix' and is the most common procedure in environmental studies where input variables are often of different types and scales.

In our December newsletter I will discuss a second method for plotting multivariate data, Nonmetric Multidimensional Scaling or NMDS. Though NMDS has a similar objective to a PCA biplot it is drawn using a different method that incorporates the information in all k variables to determine positions on the plot. After reading December's newsletter you should be better able to decide which method, NMDS or a PCA biplot, better meets your requirements for displaying in two dimensions at least some of the information in many more dimensions.

Reference:
Warwick, R.M., 1971, *Nematode associations in the Exe estuary*: Journal of the Marine Biological Association of the United Kingdom, v. 51, no. 2, p. 439–454, https://doi.org/10.1017/S0025315400031908.

C. Free Resources

There are three sources of free information you can find associated with our website practicalstats.com and our Online Training Site, practicalstats.teachable.com . These resources may help you in your use of statistical methods.

a) The Newsletter Archive at practicalstats.com

Located at https://www.practicalstats.com/news/archive.html is the archive of all our newsletters like this one going back to ancient days. Each has some information you may find useful. You can click on topics at the top of the archive to shorten the list to only those addressing that topic.

b) Videos at our Online Training Center, https://practicalstats.teachable.com/

There are now 9 free videos discussing how to analyze data with nondetects at our Training Center. Look for the <1 to <9 logos. There are also 3 videos discussing methods in environmental statistics there, and three more on our Videos page at practicalstats.com. See the entire list at https://www.practicalstats.com/training/webinar.html
and then view one or more to get a taste of the information that is available in one of our online courses.

c) 2020 Textbook

The new 2020 version of the classic textbook *Statistical Methods in Water Resources* is available as a free download from the USGS at https://pubs.er.usgs.gov/publication/tm4A3 . It serves as the textbook for our *Applied Environmental Statistics* courses. A hardback book version is also available from the USGS for about $25, including shipping within the US. See https://store.usgs.gov/product/533012 . The USGS will ship the book to addresses outside the United States for an additional charge.

'Til next time,

Dennis Helsel
ask@practicalstats.com
Practical Stats LLC
-- Make sense of your data