Practical Stats Newsletter for March 2013

Subscribe and Unsubscribe: http://practicalstats.com/news/
Download any of our past newsletters:
http://www.practicalstats.com/news/bydate.html

In this newsletter:
1. Upcoming Events
2. Cluster Analysis: a series of techniques
3. Discount on Course Registration!

1. Upcoming Events

In-person courses
Untangling Multivariate Relationships
August 13-14, 2013
Allmendinger Center, Washington State Univ.
Puyallup, WA 98371

Time Series Methods for Frequently Collected Data
August 15-16, 2013
Allmendinger Center, Washington State Univ.
Puyallup, WA 98371

Webinars:
Monday March 18. Hypothesis Tests: How They Work & Paired Data Tests. $250
Second in the Applied Environmental Statistics series. You can still register for all 6 before March 18th and save $200 (1st webinar will be available as a recording)

Monday March 25. Invasive Data: Why Not Substitute ½ the Detection Limit? $50
"Substitution is not neutral. It produces invasive data, choking off the information in measured data. Two better methods for dealing with nondetects are introduced."
Or register for the series of 4 webinars on handling nondetect data and save $100

To register and for more information on all of our courses and webinars, see our Training page at http://www.practicalstats.com/training/

2. Cluster Analysis: a series of techniques

One of the most popular methods from our Untangling Multivariate Relationships course is cluster analysis. Starting with a group of observations, analysis of similarities and differences along a suite of measurement axes (chemical concentrations, physical attributes, species counts) results in aggregation of individual observations into self-similar groups or clusters. Observations within each cluster are more similar to others in the cluster than they are to those outside the cluster. Or at least that is the goal.

Some persons believe cluster analysis is akin to a hypothesis test – plug in the data and get back 'the truth' in a unique result. It is actually very far from that! Cluster analysis is more similar to how you might choose to file documents into folders on your computer. You could choose to file them by project, all files created for the same project over several years put into one folder. Or choose to file them by month and year, with documents for all purposes completed in January 2012 filed in that month's folder. You could choose monthly folders, so 12 folders per year, or quarterly folders. The number of clusters (folders) and method of classification is up to you, chosen to best suit your objectives.

Cluster analysis returns a series of classifications, the best 2 clusters, best 3 clusters, 4 clusters, all the way up until each observation is shown in its own 'cluster'. You choose how many clusters you think are present. You also choose the variables going into the analysis, the measure of similarity or difference (distance) between observations to use, and the mechanism for combining observations together into groups. Choosing different variables, distance measures, or combining mechanisms all change the resulting clusters you get back. If you are not familiar with how these choices affect the outcome, you may not get back what you expect or believe.

For example, Figure 1 shows the result of cluster analysis on two groups of data.

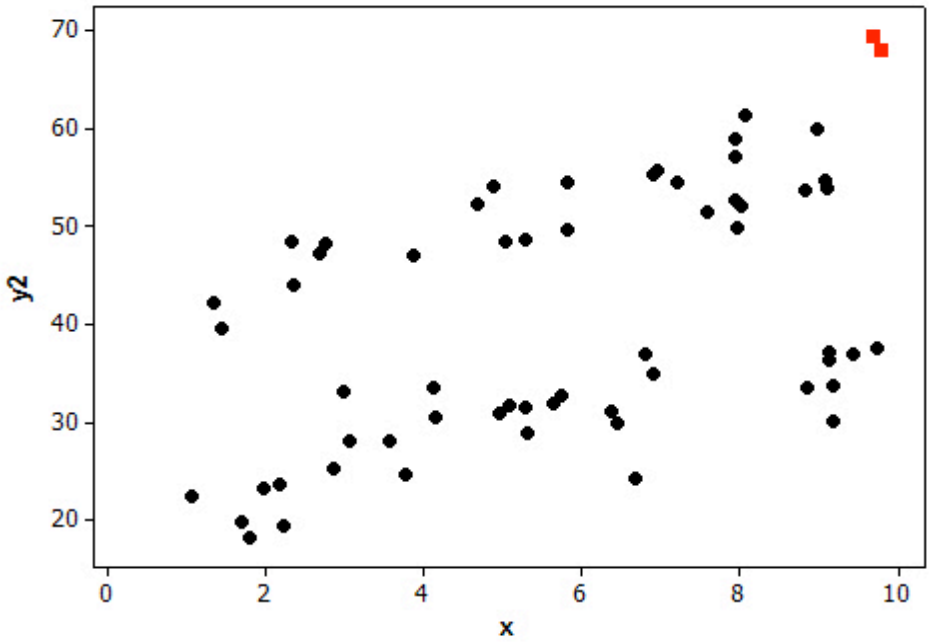

Figure 1. Cluster analysis of two groups using single linkage and Euclidean distance

The data form a linear pattern with a positive slope, one group above the other. The combining mechanism used for clustering was single linkage, which tends to split off end-members from a gradient such as the one here. Stating that there should be two clusters results in one formed by two (red) observations at the top right, leaving the all

the remaining data in the second cluster. If you knew how this combining mechanism worked, you probably would not have applied it to data representing a gradient of values.

As a second try, Figure 2 shows clustering by centroid linkage designed to look for and form spherical clusters. The resulting two clusters don't look anything like a gradient. Like using a hammer to drive a screw into the wall, using an incorrect tool to do a job can make a real mess of things.

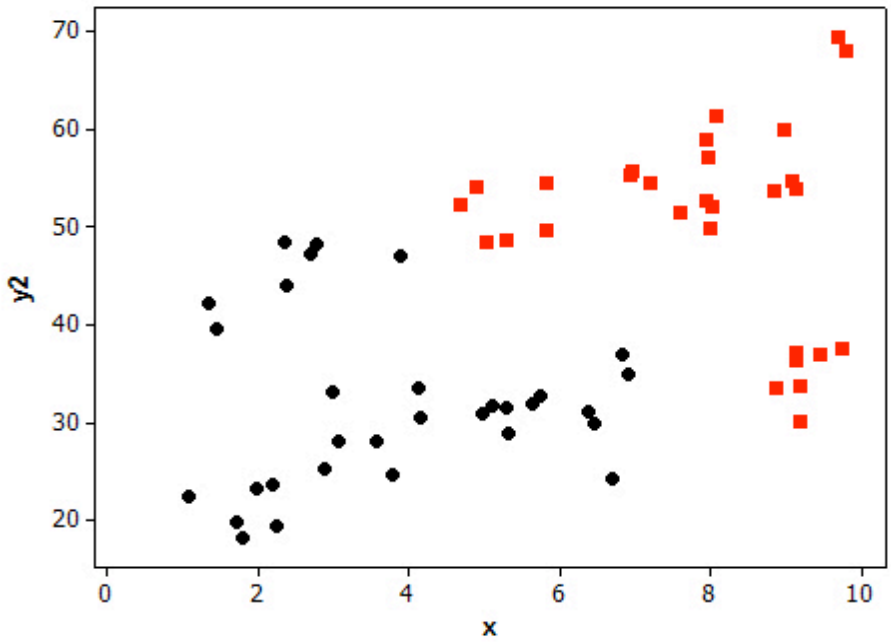

Figure 2. Cluster analysis of two groups using centroid linkage and Manhattan distance

Yet a third linkage mechanism results in clusters showing the original data pattern we started with (Figure 3). There are industry-standard practices, and this third set of linkage and distance measures is one of them. It is not as hopeless as the common accusation against multivariate methods that you can "get whatever you want to get". But users must have an understanding of what each type of linkage mechanism and distance measure was actually intended to do.

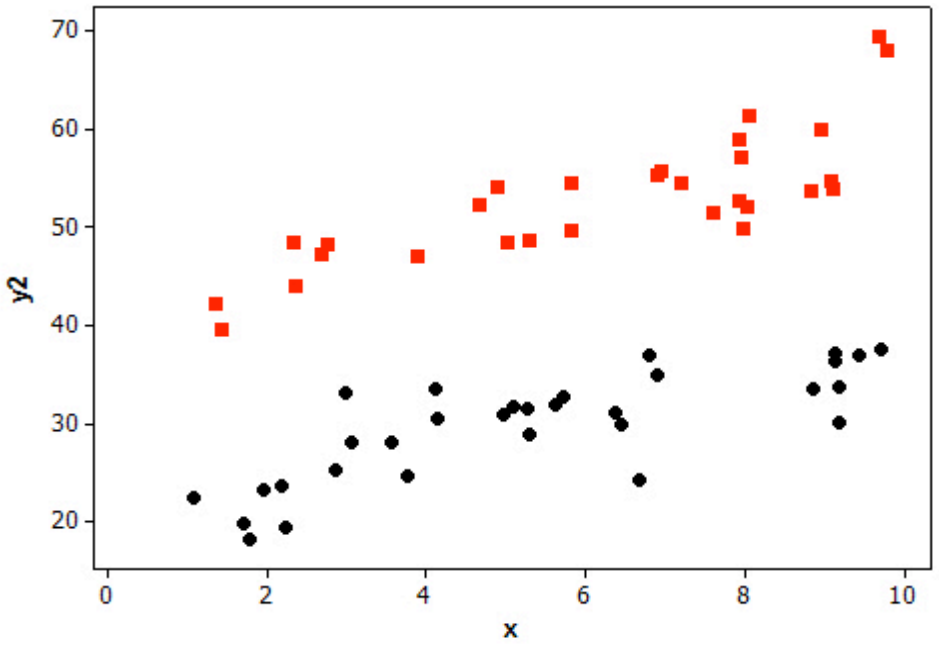

Figure 3. Cluster analysis of two groups using Ward's linkage and Euclidean distance

Like many multivariate methods, cluster analysis is an exploratory technique rather than one that packages and dispenses "the truth". It enables you to see logical patterns in your data. However, those patterns come only after you make several important choices in methodology. Knowing which tool does what is very important. It is why we spend a considerable part of our Untangling Multivariate Relationships course covering the attributes of each method, why you would use it, and when. There are interesting new developments as well, including a nonparametric test for discerning how many clusters are indicated by the data, rather than leaving this totally to your intuition.

3. Discount on Course Registration!

We publish a newsletter of tips on environmental statistics -- this pdf is the March newsletter, for example. Subscribing saves us a lot of time NOT answering email questions about "when is the next course/webinar being held?" So to say thank you in a Practical way, when you subscribe to our newsletter at
http://practicalstats.com/news/
you'll receive a discount code for 10% off registration costs to either our Untangling Multivariate Relationships or Time Series Methods for Frequently-Collected Data course this coming August.

This offer is also available to friends and co-workers of yours who sign up. Point them to our Newsletter page to sign up, and the discount is theirs also!

'Til next time,

Practical Stats (Dennis Helsel)
-- Make sense of your data