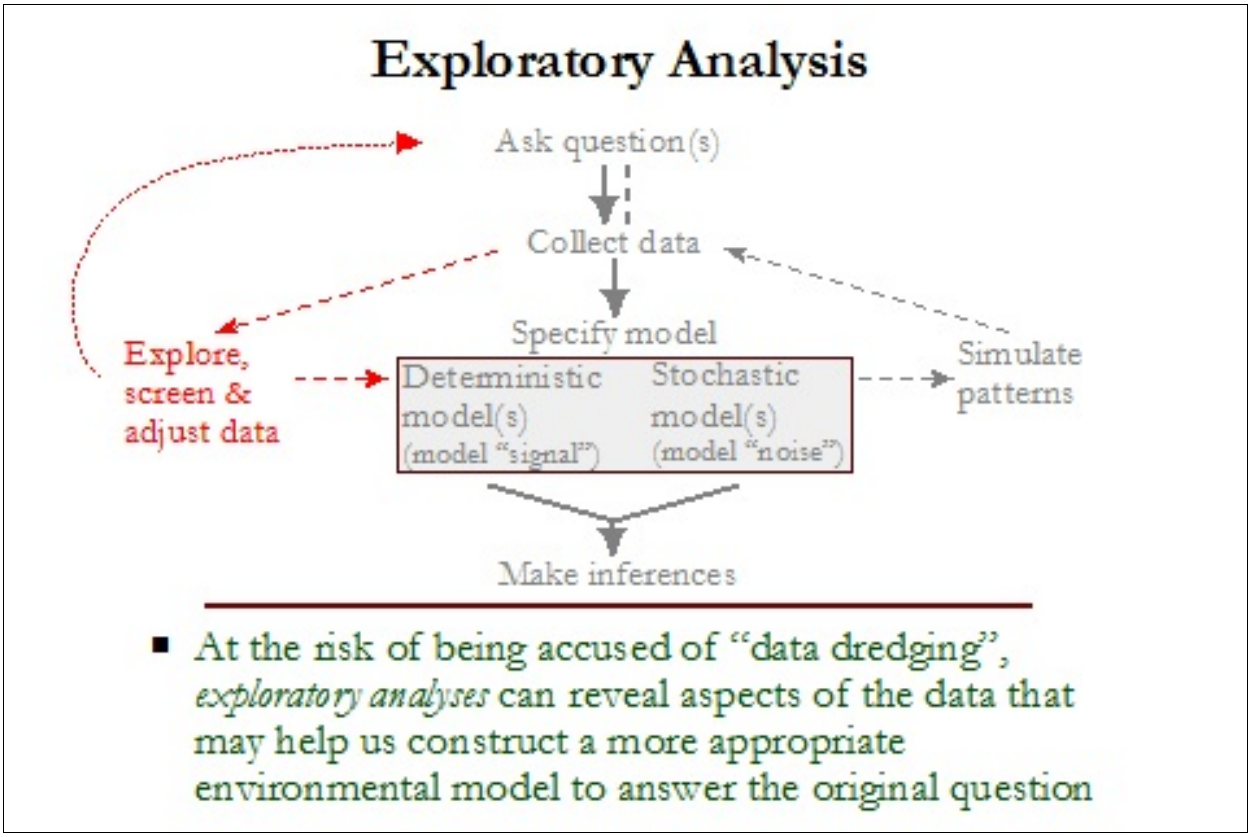
# Analysis of Environmental Data

## Chapter 3. Conceptual Foundations:

*Data Exploration, Screening & Adjustments*




## 1. Purpose of data exploration, screening & adjustments

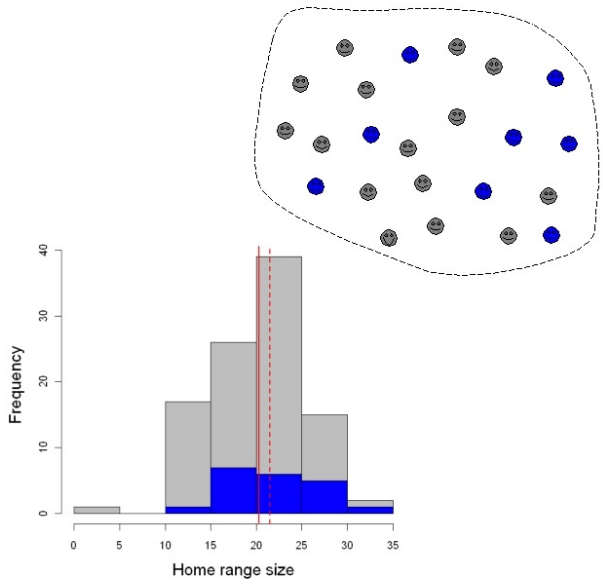
One of the basic tensions in all data analysis and modeling is how much you have all your questions framed before you begin to look at your data. In the classical statistical framework, you are suppose to have all your hypotheses laid out in advance and not stray from that course during the analysis. Allowing your data to suggest new statistical tests raises the risk of "fishing expeditions" or "data-dredging" – indiscriminate scanning of the data for patterns. But this philosophy may be too strict for environmental scientists. Unexpected patterns in the data can inspire you to ask new questions, and it is foolish not to explore your hard-earned data in this regard. In addition, exploratory analyses can reveal aspects of the data that may help you construct a more appropriate environmental model to answer the original question. I see no particular harm in letting the data guide you to a better model, as long as you recognize the risk of detecting patterns that are not real and seek to confirm the findings with subsequent study. Moreover, it is always prudent to screen your data for problems before undertaking a sophisticated statistical model. In particular, you may have missing data values which may cause problems later if they are not dealt with up front. Some variables may not contain sufficient information content to warrant including them in the analysis and you want to identify those variables and remove them early on. There may be a need to transform and/or standardize variables to put them on equal footing in the analysis or better meet statistical assumptions or change the data to better reflect the environmental question. And lastly, there is always a need to screen your data for extreme values, or "outliers" which can exert undue pull on the analysis.

# Exploratory Analysis... common parameters & statistics

## Parameters versus Statistics

- *Parameters*... measured characteristics of the population
- *Statistics*... measured characteristics of the sample

Statistical inference involves estimating population parameters from sample statistics




## 2. Common parameters and statistics

*2.1 Parameters and statistics*

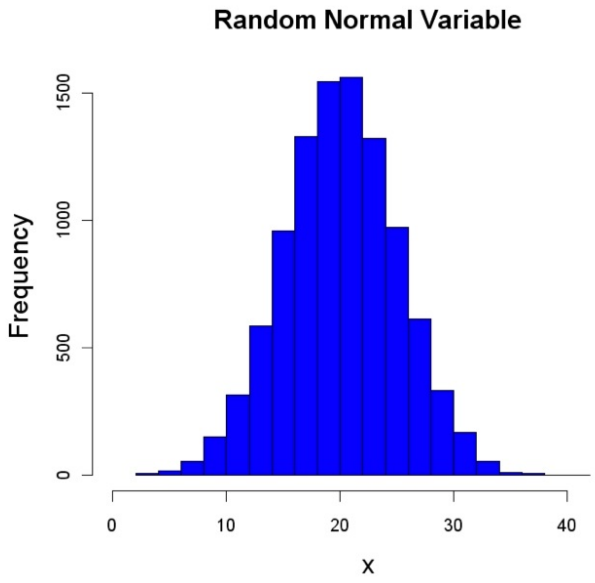
There are many common parameters (population) and statistics (sample) that are commonly used to describe data patterns; it behooves you to become very familiar with these as a means of describing your data and screening for problems before you attempt to analyze a more complex statistical model. First, it is important to understand the difference between a parameter and a statistic in the formal sense.

- *Parameters*... measured characteristics of the population, usually unknown and/or unknowable yet the thing we are most interested in knowing.
- *Statistics*... measured characteristics of the sample, which we typically use to estimate population parameters that we cannot measure directly. Statistics are the basis for all of statistical inference, since to infer is to draw conclusions about a population from a sample.

# Exploratory Analysis... common parameters & statistics

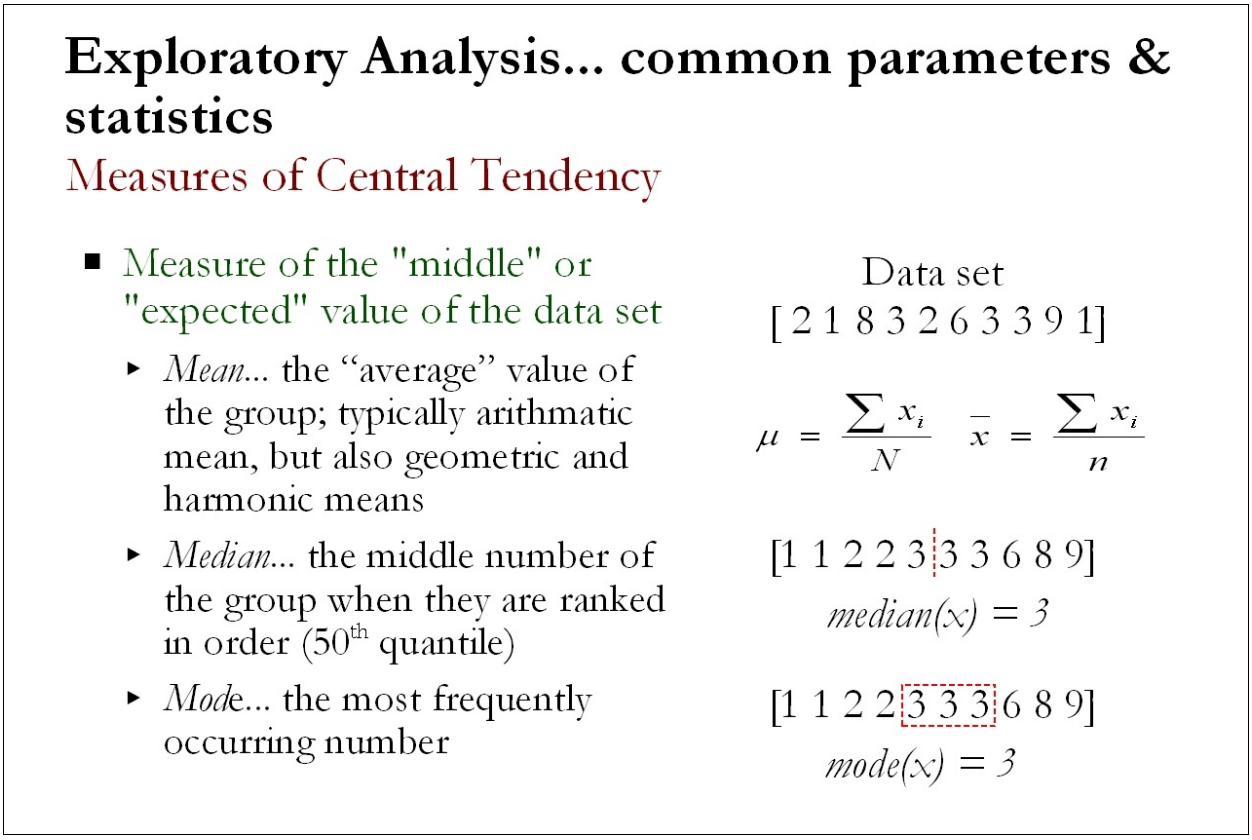
## The "Normal" Distribution

- The "normal" distribution is frequently used as a reference framework (and is the basis for most classical statistics)
- Characteristic symmetrical "bell-shaped" distribution




*2.2 The "normal" distribution*

Before describing some of the more common parameters and statistics, however, it is necessary to introduce the "normal" distribution because it is frequently used as a reference framework for describing data characteristics and is the basis for most classical statistics. We will describe this distribution more formally in a later section, but for now, suffice it to say that a normal distribution describes a data set that exhibits a symmetrical "bell- shaped" frequency distribution. That is, a collection of values that concentrate around a single central tendency (the average value) and trail off in both directions at the same rate.




*2.3 Measures of central tendency*

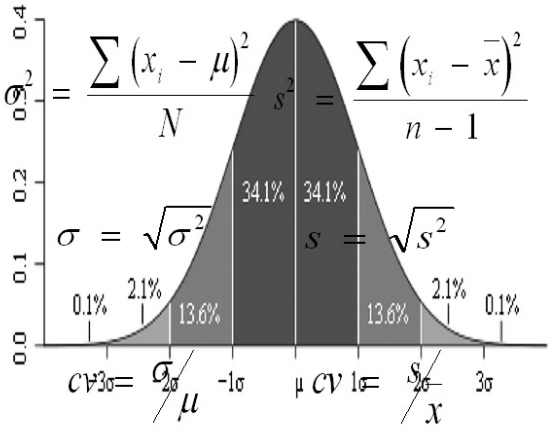
Measures of central tendency measure of the "middle" or "expected" value of the data set and are used almost universally to describe sample characteristics and as the basis for statistical inference. There are many different measures of central tendency, but some of the most common are as follows:

- *Mean*... the "average" value of the group; typically the arithmetic mean is used in environmental science, but some circumstances warrant using the geometric or harmonic means.
- *Median*... the middle number of the group when they are ranked in order (50th quantile).
- *Mode*... the most frequently occurring number.

## Exploratory Analysis... common parameters & statistics

### Measures of Spread

- Measure of the dispersion of the data set or how spread out the data is
  - *Variance*... mean squared deviation from the mean or expected value
  - *Standard deviation*... root mean squared deviation from the mean or expected value
  - *Coefficient of variation*... normalized measure of spread; relative standard deviation




*2.4 Measures of spread*

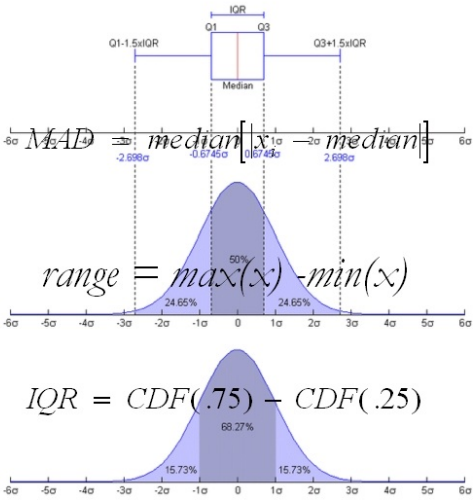
Measures of spread measure the dispersion of the data set or how spread out the data is and are used almost universally to describe sample characteristics and as the basis for statistical inference. There are many different measures of spread, but some of the most common based on deviations from the mean, or the spread about the mean, are as follows:

- *Variance*... mean squared deviation from the mean or expected value. Note, the units are squared so they are somewhat uninterpretable.
- *Standard deviation*... root mean squared deviation from the mean or expected value; i.e., the square root of the variance. The standard deviation has special meaning for normally distributed variables, because the mean ± 1 standard deviation captures approximately 68% of the values, ± 2 standard deviations captures approximately 95% of the values, and ± 3 standard deviation captures more than 99% of the values. Note, the standard deviation is in the same units as the measurement variable, which are therefore interpretable, unlike the variance which is in squared units.
- *Coefficient of variation*... normalized measure of spread, defined as the standard deviation divided by the mean (often multiplied by 100 to express as a percentage). The coefficient of variation allows us to compare the spread for variables measured on different scales.

# Exploratory Analysis… common parameters & statistics

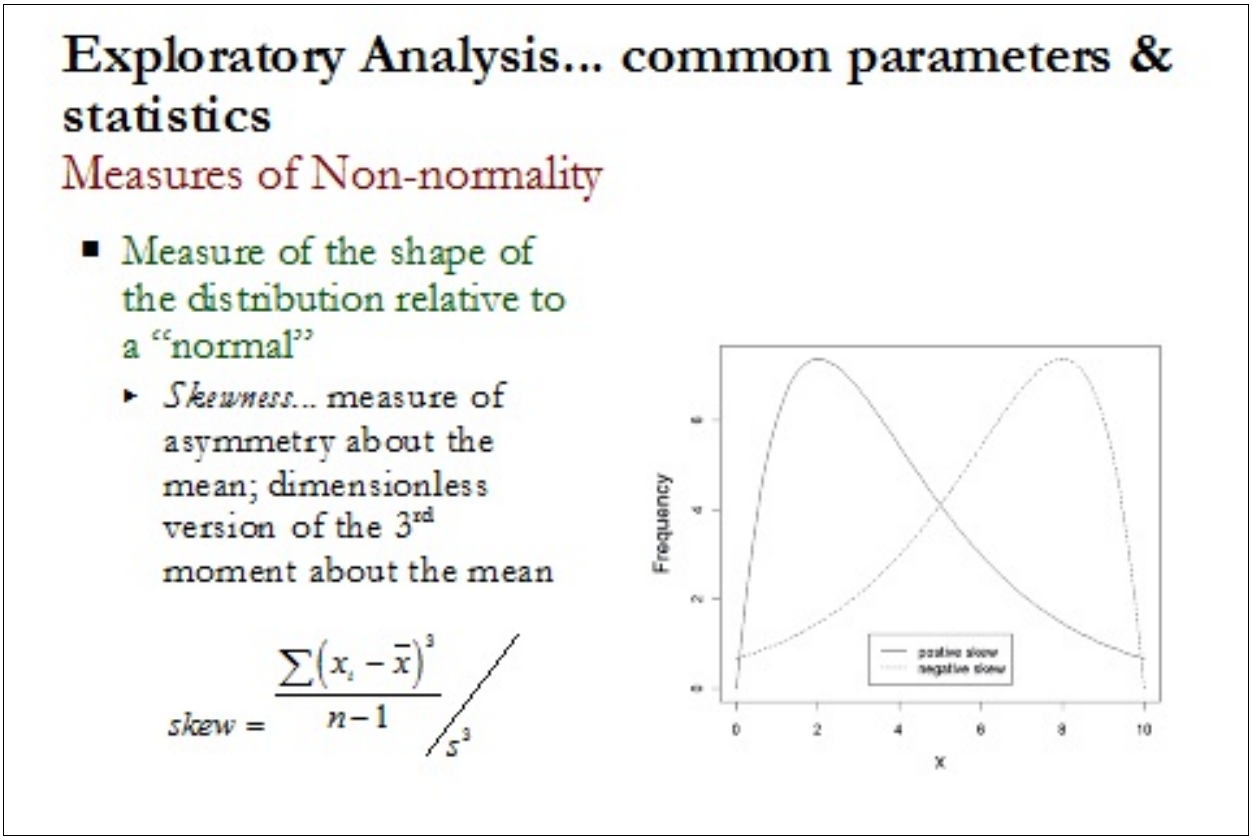
## Measures of Spread

- Measure of the dispersion of the data set or how spread out the data is
  - *Median absolute deviation*… median absolute deviation from the median
  - *Range*… absolute range of values (from min to max)
  - *Interquartile range*… range between the 25th and 75th quantiles of the data




Not all measures of spread are based on deviations from the mean. Some measures refer to deviation from the median or describe the absolute range of values or the range between certain quantiles of the data.. For example:

- *Median absolute deviation*… median absolute deviation from the median (MAD).
- *Range*… absolute range of values (from min to max).
- *Interquartile range*… range between the 25th and 75th quantiles of the data (IQR). Note, the IQR ± 1.5xIQR captures roughly 99% of the distribution and is roughly equivalent to the mean ± 3 standard deviations. Moreover, the IQR and the IQR ± 1.5xIQR is the basis for a box-and-whisker plot (see later).




*2.5 Measures of non-normality*

Measures of non-normality measure the shape of the distribution relative to a "normal" distribution; i.e., deviation from the symmetric bell-shaped distribution. The most common statistics are skewness and kurtosis:

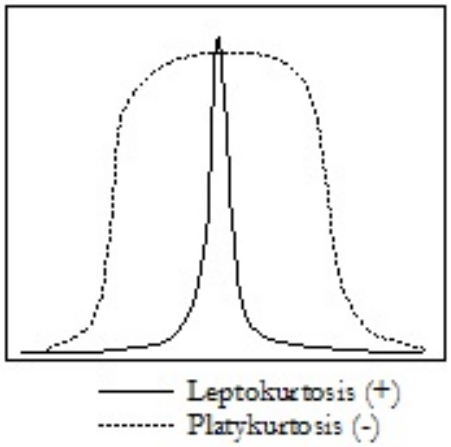
- *Skewness*... measure of asymmetry about the mean; it is a dimensionless version of the 3rd moment about the mean. Notice the similarity to the variance, except that the deviations from the mean are cubed instead of squared. The denominator is the standard deviation cubed, which is normalizing constant and makes the statistic dimensionless, since the units cancel each other out. Defined in this manner, skewness = 0 for a normal distribution. A positively skewed distribution (also called right skewed) is a distribution with a longer right-side tail, which is prevalent with environmental data.

## Exploratory Analysis... common parameters & statistics

### Measures of Non-normality

- Measure of the shape of the distribution relative to a "normal"
  - *Kurtosis*... measure of peakedness, or flat-toppedness, of a distribution; dimensionless version of the 4th moment about the mean

$$kurtosis = \left( \frac{\sum (x_i - \bar{x})^4}{(n-1)s^4} \right) - 3$$

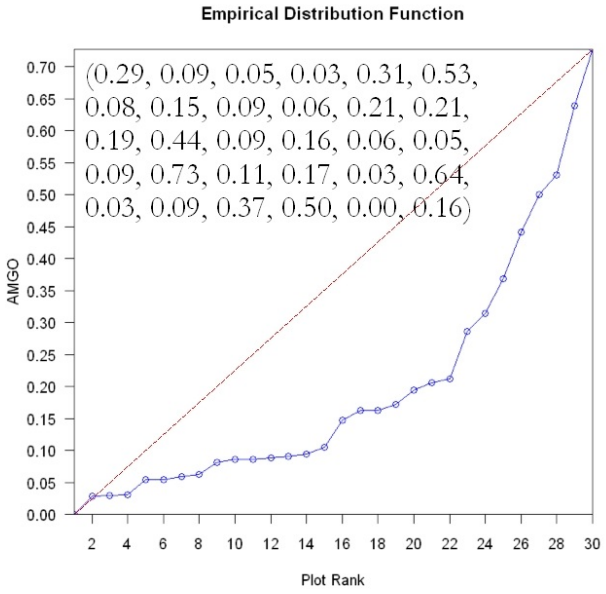



- *Kurtosis*... measure of peakedness, or flat-toppedness, of a distribution; it is a dimensionless version of the 4th moment about the mean. Again, notice the similarity to the variance, except that the deviations from the mean are raised to the 4th power instead of the 2nd power. The denominator is the standard deviation raised to the 4th power, which is normalizing constant and makes the statistic dimensionless, since the units cancel each other out. A normal distribution has a value of 3, so often this is subtracted so that kurtosis = 0 for a normal distribution. Defined in this way, a positive kurtosis is more peaked than normal and is known as a leptokurtotic distribution, whereas a negative kurtosis is more flat topped than normal and is known as a platykurtotic distribution. To help remember the distribution, remember that plat rhymes with flat, plat is short for plateau, and plat is short for platypus which have a flat square-tipped tail.

## Exploratory Analysis... single variable plots

### Empirical Distribution Plot

- Graphical display of the *rank order* distribution of increasing values of the variable
  - A *uniform* distribution of values will have points that fall on a perfect diagonal straight line
  - A *normal* distribution will have a sigmoidal shape




## 3. Single variable plots

While the common parameters and statistics described above are quite useful for describing variables and their distributions, most people find graphical summaries more compelling and informative. There are myriad types of plots for single variables – too many to cover here. However, there are several common graphical plots that are nearly universally used for continuous variables, so it behooves us to understand these at a minimum.

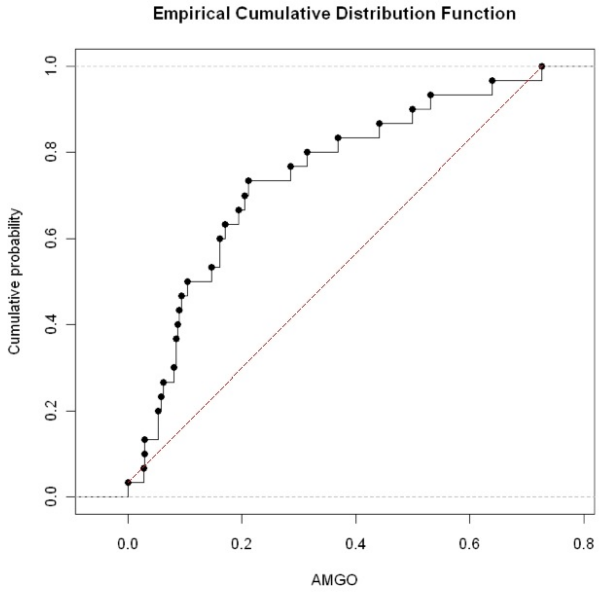
### *3.1 Empirical distribution function*

The empirical distribution function (EDF) is a simple rank order distribution of increasing values of the variable. A variable with a perfectly uniform distribution of values within its range (minimum to maximum), so that no one value is more common than another, will have points that fall on a perfect diagonal straight line. Deviations from the diagonal indicate non-uniformity. A normally distributed variable will have a sigmoidal shape. Deviations from these reference lines can be quite useful in quickly revealing departures from these common distributions.

## Exploratory Analysis... single variable plots

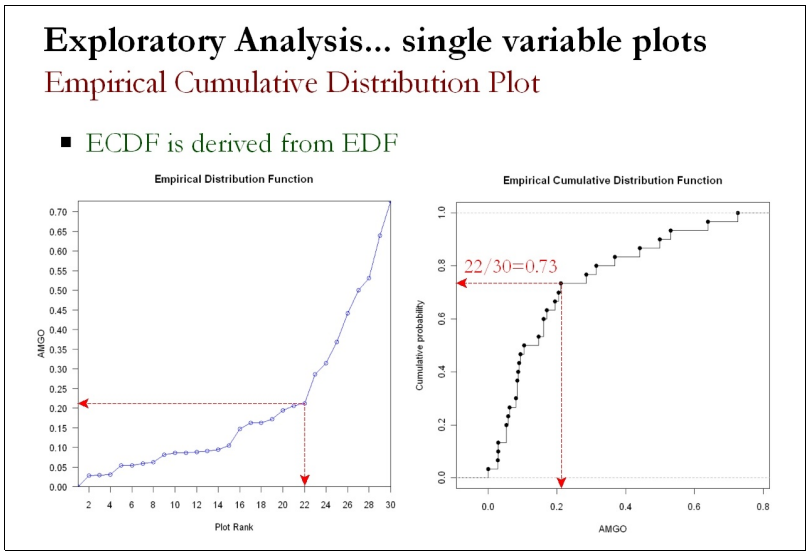
### Empirical Cumulative Distribution Plot

- Graphical display of the probability of values falling below any given value of *x* on any observation
  - A *uniform* distribution of values will have points that fall on a perfect diagonal straight line
  - A *normal* distribution will have a sigmoidal shape




*3.2 Empirical cumulative distribution function*

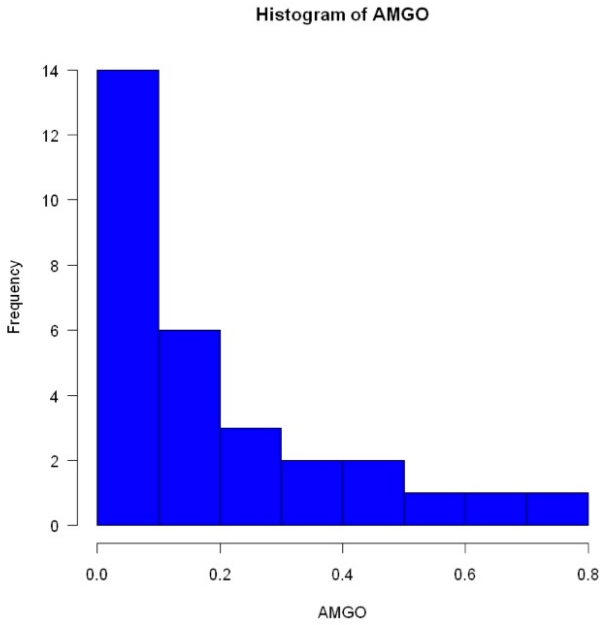
The empirical cumulative distribution function (ECDF or just CDF) is derived from the EDF and gives the probability, or proportion, of values (the y-axis) falling below any given value of x (the x-axis). The ECDF is read as follows. Start with any value of the *x* variable and draw a vertical line upwards from that point on the *x* axis until the line intersects the ECDF. From that point draw a horizontal line to the left until the line intersects the *y* axis and read the probability. That probability is the probability of observing the specified value of the *x* variable or a lower value. Note, the "cumulative" probability means that the probability refers to a value equal to or less then the specified *x* value.



## Exploratory Analysis... single variable plots

### Histogram

- Graphical display of tabulated frequencies (or probabilities), shown as bars
- Shows what proportion of cases fall into each of several adjacent non-overlapping categories
- A way of binning the data



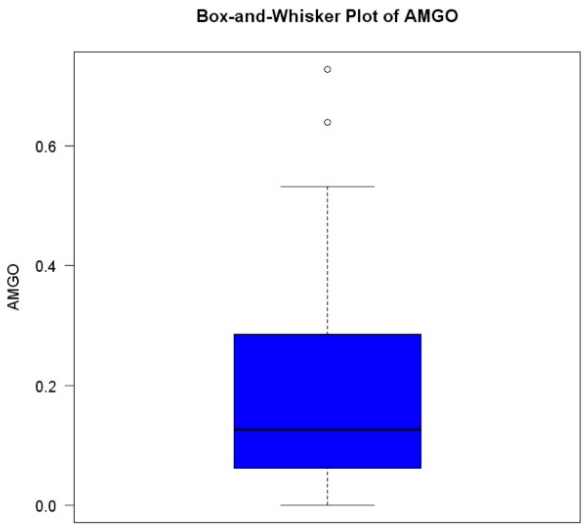

*3.3 Histogram*

A histogram is a graphical display of tabulated frequencies (or probabilities), shown as bars; it shows what proportion of cases fall into each of several adjacent non-overlapping categories; and it is a way of binning the data. Histograms are extremely useful for quickly visualizing the distribution of values in a variable. As environmental data are often highly skewed, a histogram will readily reveal that skew. Moreover, extreme values in the distribution (i.e., potential outliers), which we will discuss later, often show up as isolated bars on the tail of the distribution. In addition, for many kinds of data, such as species abundance variables, pay attention to the level of quantitative information present (i.e., whether there is a range of abundances or whether the principal signature is one of presence versus absence) since this may determine the need for a binary transformation (discussed later).

# Exploratory Analysis... single variable plots

## Box and Whisker Plot

- Graphical display of the spread of a variable
  - *Solid line* depicts the median (50th quantile)
  - *Box* depicts the inter-quartile (25-75th quantiles) range
  - *Whiskers* depict the range of the data up to 1.5xIQR
  - *Isolated points* depict "extreme" values



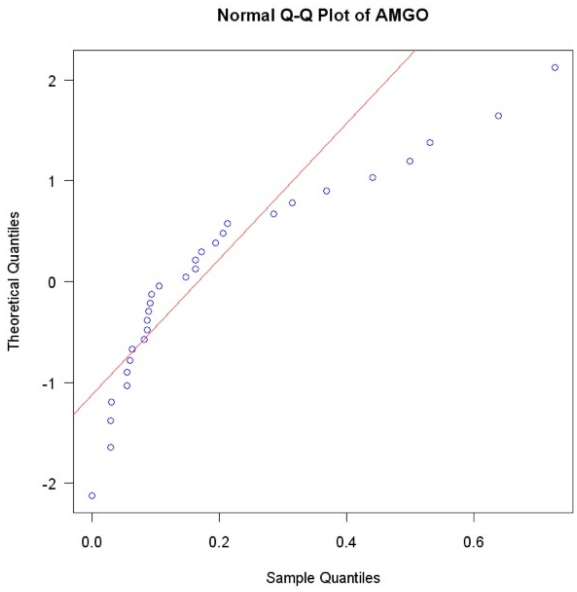

*3.4 Box-and-whisker plot*

An alternative way to examine the distribution of a variable is with a box-and-whisker plot. Box and whisker plots can depict the skewness of a distribution quite nicely and also can be used to identify extreme observations. The central box shows the data between the 'hinges' (roughly quartiles), with the median represented by a solid line. 'Whiskers' go out to the extremes of the data, and very extreme points (defined as samples that are by default farther than 1.5 times the inter-quartile range from the box) are shown by themselves.

## Exploratory Analysis... single variable plots

### Normal Quantile Plot

- Graphical display of the *sample* quantiles on the x axis against the *theoretical* quantiles from a normal distribution of the same sample size on the y axis
- A perfectly *normal* distribution of values will have points that fall on a perfect diagonal straight line




*3.5 Normal quantile-quantile plot*

Another useful way of examining the distribution of each variable is to compare the empirical cumulative distribution function (ECDF) to the expected ECDF for a normal distribution. A normal quantile-quantile (or qqnorm) plot does just this. The qqnorm plot depicts the sample quantiles on the *x* axis against the theoretical quantiles from a normal distribution of the same sample size on the *y* axis. If the data are from a perfectly normal distribution, the data will lie on a diagonal straight line. Departures from the diagonal indicate deviations from a normal distribution. Skewed distributions show up nicely as deviations from the line at the tails.

# Exploratory Analysis... single variable plots

## Normal Quantile Plot

| AMGO | rAMGO | qAMGO | q2AMGO | zAMGO | qNorm | q2Norm |
|---|---|---|---|---|---|---|
| 0 | 1 | 0.03 | 0.02 | -1.04 | -1.83 | -2.13 |
| 0.03 | 2 | 0.07 | 0.05 | -0.89 | -1.5 | -1.64 |
| 0.03 | 3 | 0.1 | 0.08 | -0.89 | -1.28 | -1.38 |
| 0.03 | 4 | 0.13 | 0.12 | -0.88 | -1.11 | -1.19 |
| 0.05 | 5.5 | 0.18 | 0.15 | -0.76 | -0.9 | -1.04 |
| 0.05 | 5.5 | 0.18 | 0.18 | -0.76 | -0.9 | -0.9 |
| 0.06 | 7 | 0.23 | 0.22 | -0.73 | -0.73 | -0.78 |
| 0.06 | 8 | 0.27 | 0.25 | -0.72 | -0.62 | -0.67 |
| 0.08 | 9 | 0.3 | 0.28 | -0.62 | -0.52 | -0.57 |
| 0.09 | 10.5 | 0.35 | 0.32 | -0.59 | -0.39 | -0.48 |
| 0.09 | 10.5 | 0.35 | 0.35 | -0.59 | -0.39 | -0.39 |
| 0.09 | 12 | 0.4 | 0.38 | -0.58 | -0.25 | -0.3 |
| 0.09 | 13 | 0.43 | 0.42 | -0.57 | -0.17 | -0.21 |
| 0.09 | 14 | 0.47 | 0.45 | -0.55 | -0.08 | -0.13 |
| 0.11 | 15 | 0.5 | 0.48 | -0.49 | 0 | -0.04 |
| 0.15 | 16 | 0.53 | 0.52 | -0.28 | 0.08 | 0.04 |
| 0.16 | 17.5 | 0.58 | 0.55 | -0.2 | 0.21 | 0.13 |
| 0.16 | 17.5 | 0.58 | 0.58 | -0.2 | 0.21 | 0.21 |
| 0.17 | 19 | 0.63 | 0.62 | -0.15 | 0.34 | 0.3 |
| 0.19 | 20 | 0.67 | 0.65 | -0.03 | 0.43 | 0.39 |
| 0.21 | 21 | 0.7 | 0.68 | 0.03 | 0.52 | 0.48 |
| 0.21 | 22 | 0.73 | 0.72 | 0.06 | 0.62 | 0.57 |
| 0.29 | 23 | 0.77 | 0.75 | 0.44 | 0.73 | 0.67 |
| 0.31 | 24 | 0.8 | 0.78 | 0.59 | 0.84 | 0.78 |
| 0.37 | 25 | 0.83 | 0.82 | 0.87 | 0.97 | 0.9 |
| 0.44 | 26 | 0.87 | 0.85 | 1.25 | 1.11 | 1.04 |
| 0.5 | 27 | 0.9 | 0.88 | 1.55 | 1.28 | 1.19 |
| 0.53 | 28 | 0.93 | 0.92 | 1.72 | 1.5 | 1.38 |
| 0.64 | 29 | 0.97 | 0.95 | 2.27 | 1.83 | 1.64 |
| 0.73 | 30 | 1 | 0.98 | 2.73 | Inf | 2.13 |

- AMGO = raw data
- rAMGO = rank
- qAMGO = quantile
- q2AMGO = adjusted quantile; no ties and with offset as in qqnorm() ((1:30-.5)/(30+(1-.5)-.5)
- zAMGO = z-scores (x-mean(x)/sd(x))
- qNorm = quantile value of standard normal (mean=0, sd=1) for qAMGO quantiles
- q2Norm = same but for q2AMGO quantiles

The normal quantile-quantile plot is a bit confusing at first and so it warrants some additional explanation of how it is created. First, we sort the variable (AMGO in this case) from smallest to largest value and assign a rank (rAMGO) and the corresponding quantile of the data it represents (qAMGO). As it turns out, there are several different methods for computing quantiles depending on how one treats ties and the bookends (min and max) data values, but we won't go into the details of the various methods here. Importantly, computing the quantile value of a standard normal distribution (below) is somewhat problematic for one or both of the bookends. Consequently, it is conventional to adjust the quantiles by an offset (q2AMGO); e.g., given as: (1:n - a)/(n + (1-a)-a), where n is the number of observations, and a = ifelse(n <= 10, 3/8, ½)). Thus, in our case, a = 0.5, and instead of a quantile of 1 for the maximum data value, we end up with 0.98.

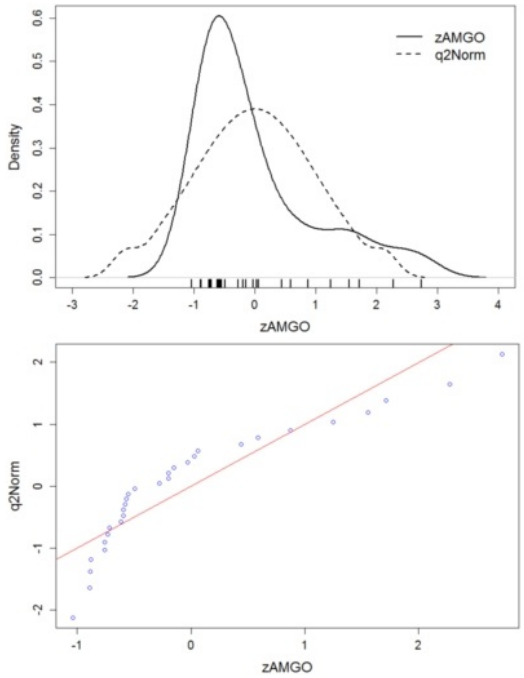
Next, for plotting purposes only, we can adjust the AMGO data values to their corresponding z-scores (zAMGO) by subtracting the mean and dividing by the standard deviation, such that the resulting values have a mean = 0 and sd = 1. Note, this does not change the shape of the distribution, only the scale of the axis.

Lastly, for either the original quantiles of AMGO (qAMGO) or the adjusted quantiles (q2AMGO), we compute the corresponding quantile values of a standard (i.e., z-scores) Normal distribution (qNorm and q2Norm, respectively). Thus, for each value of AMGO we have the corresponding quantile value of a theoretical Normal distribution with the same mean and standard deviation.

# Exploratory Analysis... single variable plots

## Normal Quantile Plot

| AMGO | rAMGO | qAMGO | q2AMGO | zAMGO | qNorm | q2Norm |
|---|---|---|---|---|---|---|
| 0 | 1 | 0.03 | 0.02 | -1.04 | -1.83 | -2.13 |
| 0.03 | 2 | 0.07 | 0.05 | -0.89 | -1.5 | -1.64 |
| 0.03 | 3 | 0.1 | 0.08 | -0.89 | -1.28 | -1.38 |
| 0.03 | 4 | 0.13 | 0.12 | -0.88 | -1.11 | -1.19 |
| 0.05 | 5.5 | 0.18 | 0.15 | -0.76 | -0.9 | -1.04 |
| 0.05 | 5.5 | 0.18 | 0.18 | -0.76 | -0.9 | -0.9 |
| 0.06 | 7 | 0.23 | 0.22 | -0.73 | -0.73 | -0.78 |
| 0.06 | 8 | 0.27 | 0.25 | -0.72 | -0.62 | -0.67 |
| 0.08 | 9 | 0.3 | 0.28 | -0.62 | -0.52 | -0.57 |
| 0.09 | 10.5 | 0.35 | 0.32 | -0.59 | -0.39 | -0.48 |
| 0.09 | 10.5 | 0.35 | 0.35 | -0.59 | -0.39 | -0.39 |
| 0.09 | 12 | 0.4 | 0.38 | -0.58 | -0.25 | -0.3 |
| 0.09 | 13 | 0.43 | 0.42 | -0.57 | -0.17 | -0.21 |
| 0.09 | 14 | 0.47 | 0.45 | -0.55 | -0.08 | -0.13 |
| 0.11 | 15 | 0.5 | 0.48 | -0.49 | 0 | -0.04 |
| 0.15 | 16 | 0.53 | 0.52 | -0.28 | 0.08 | 0.04 |
| 0.16 | 17.5 | 0.58 | 0.55 | -0.2 | 0.21 | 0.13 |
| 0.16 | 17.5 | 0.58 | 0.58 | -0.2 | 0.21 | 0.21 |
| 0.17 | 19 | 0.63 | 0.62 | -0.15 | 0.34 | 0.3 |
| 0.19 | 20 | 0.67 | 0.65 | -0.03 | 0.43 | 0.39 |
| 0.21 | 21 | 0.7 | 0.68 | 0.03 | 0.52 | 0.48 |
| 0.21 | 22 | 0.73 | 0.72 | 0.06 | 0.62 | 0.57 |
| 0.29 | 23 | 0.77 | 0.75 | 0.44 | 0.73 | 0.67 |
| 0.31 | 24 | 0.8 | 0.78 | 0.59 | 0.84 | 0.78 |
| 0.37 | 25 | 0.83 | 0.82 | 0.87 | 0.97 | 0.9 |
| 0.44 | 26 | 0.87 | 0.85 | 1.25 | 1.11 | 1.04 |
| 0.5 | 27 | 0.9 | 0.88 | 1.55 | 1.28 | 1.19 |
| 0.53 | 28 | 0.93 | 0.92 | 1.72 | 1.5 | 1.38 |
| 0.64 | 29 | 0.97 | 0.95 | 2.27 | 1.83 | 1.64 |
| 0.73 | 30 | 1 | 0.98 | 2.73 | Inf | 2.13 |




Now we are ready to produce the QQnorm plot. In the top figure shown here, we plotted kernel density curves (for now, think of these as simply smoothed histograms) for the z-scores of AMGO (zAMGO) and for the corresponding z-scores of a standard Normal distribution. Note, because we are plotting z-scores, both variables have the same mean = 0 and sd = 1. Thus, the differences in the shapes of the curves reflects differences between the empirical distribution of AMGO and a corresponding theoretical Normal distribution. As you can see, the shapes of the curves differ, with the zAMGO displaying a positive or right skew and the q2Norm displaying the expected bell shape for a Gaussian (or Normal) distribution.

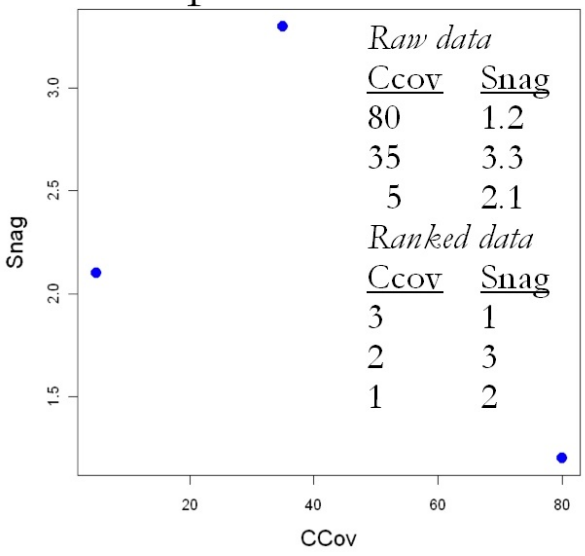
In the bottom figure, we plotted the z-scores of AMGO (zAMGO) against the corresponding quantile values for the standard Normal distribution (q2Norm). Note, here we plotted q2Norm instead of qNorm because this is the convention used in R in the qqnorm() function. Superimposed on the plot is a diagonal line through the origin (0,0) with a slope of 1. If the empirical distribution of AMGO (represented here by the corresponding z-scores) was identical to a Normal distribution (represented here by the z-scores of a Normal distribution, q2Norm), the points would all fall on the diagonal line. In this case, however, due to the right-skewed distribution of AMGO, the points fall off the diagonal line. Note, although we plotted the z-scores of AMGO in the plot shown here, we could easily substitute the z-scores with the raw data values for the x-axis, as is the convention in the qqnorm() in R. Nothing would change except the scale of the x-axis, but the interpretation is the same.

# Exploratory Analysis… measures of association

## Covariance and Correlation

- *Covariance* is a measure of association between two variables (i.e., how well do they covary)
- *Correlation* is a normalized measured of covariation
  - *Pearson's r* = covariance of two *z-score* standardized variables
  - *Spearman's rho* = covariance of two *rank* transformed variables

Example:




*Raw data*

| Ccov | Snag |
|---|---|
| 80 | 1.2 |
| 35 | 3.3 |
| 5 | 2.1 |

*Ranked data*

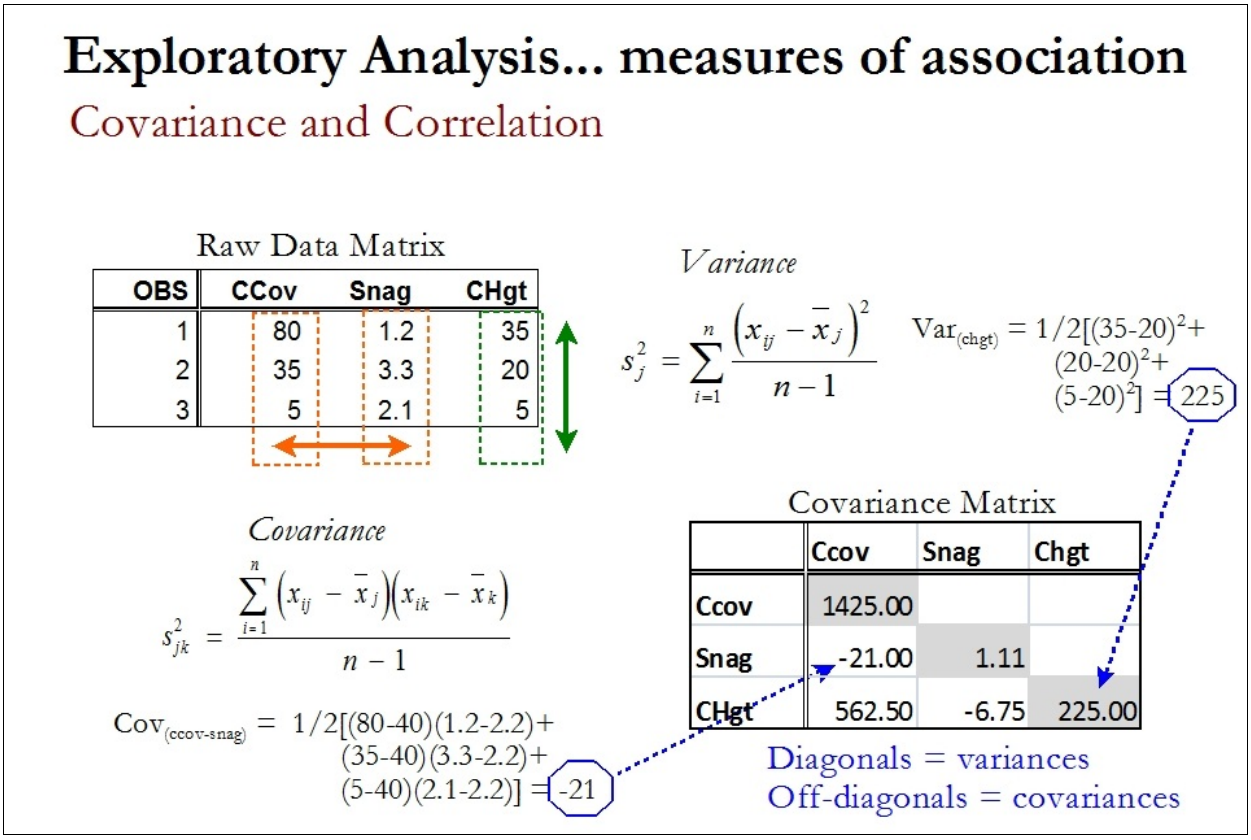
| Ccov | Snag |
|---|---|
| 3 | 1 |
| 2 | 3 |
| 1 | 2 |

## 4. Measures of association

Before considering a formal statistical analysis involving multiple variables, it is always useful to examine the nature of the relationships between pairs of variables, including both dependent and interdependent relationships. These relationships can be critically important in determining the form of the statistical relationship between the independent and dependent variables and evaluating the underlying assumptions (e.g., linearity, multicolinearity) of the model to be employed.

The most basic measure of association between two variables is known as *covariance* and its normalized version is known as *correlation*. There are different correlation coefficients, but the two most commonly used in environmental studies are as follows:

- *Pearson's r* = covariance of two *z-score* standardized variables. Z-score standardized variables are variables in which the data set is centered on zero and the spread is scaled such that the variance and standard deviation equal 1 (more on this later).
- *Spearman's rho* = covariance of two *rank* transformed variables. In this case, the data set is first transformed to ranks and then the covariance is calculated. Spearman's *rho* is a more appropriate measure of association than Pearson's *r* for non-linear associations.




Given the importance of covariance and correlation in statistical modeling, it is worth spending a little time working through an example, especially so that we can see the relationship between covariance and correlation.

The raw data matrix shown here contains 3 observations (rows) with 3 variables measured at each site: Canopy cover (CCov), snag density (Snag), and canopy height (CHgt). The sample variance is given for each variable (column) by calculating the average squared deviation from the mean. Note, the customary n-1 in the denominator to adjust for the sample bias. Note also that the variance is calculated separately for each variable. The sample covariance is given for each pair of variables and shown in detail for the covariance between CCov and Snag. Note the similarity between the formula for variance and covariance. Covariance is calculated much the same way as variance, except that the deviations from the mean of each variable are multiplied instead of squaring the deviations from the single variable. Thus, the units are again squared and meaningless. The covariance matrix (also called the variance-covariance matrix) is simply a square symmetrical matrix of variances and covariances, with the variances along the diagonal and the covariances in the off-diagonal positions. Since the matrix is symmetric, only the lower triangle is shown.

# Exploratory Analysis... measures of association

## Covariance and Correlation

Raw Data Matrix

| OBS | CCov | Snag | CHgt |
|---|---|---|---|
| 1 | 80 | 1.2 | 35 |
| 2 | 35 | 3.3 | 20 |
| 3 | 5 | 2.1 | 5 |

*Correlation (r)*

$$\frac{n\sum\left(x_{ij}x_{ik}\right)-\sum x_{ij}\sum x_{ik}}{\sqrt{\left[n\sum x_{ij}^2-\left(\sum x_{ij}\right)^2\right]\left[n\sum x_{ik}^2-\left(\sum x_{ik}\right)^2\right]}}$$

$$Cor_{(ccov\text{-}ccov)} = \frac{3[(80)(80)+(35)(35)+(5)(5)]-[(120)(120)]}{\{[3(80^2+35^2+5^2)-(120^2)]\,[3(80^2+35^2+5^2)-(120^2)]\}1/2} = 1.000$$

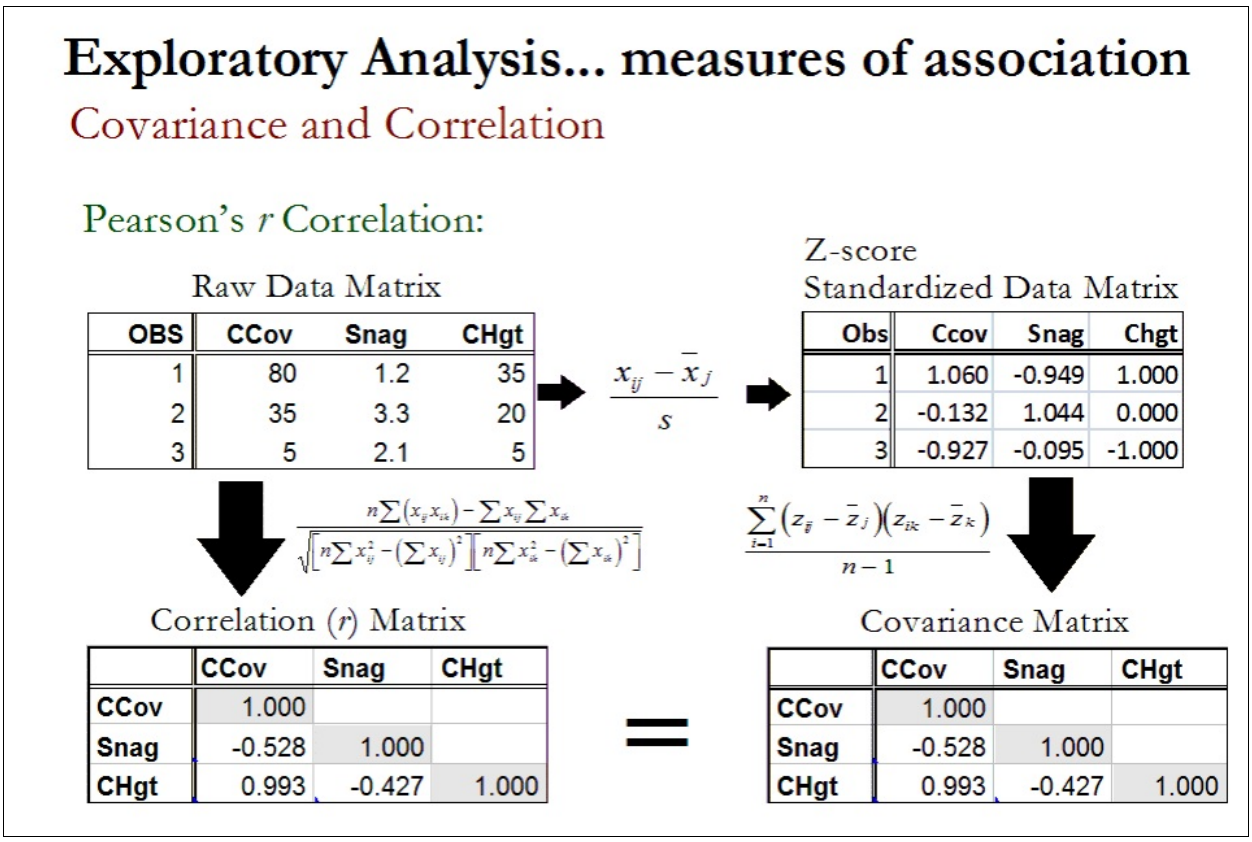
$$Cor_{(ccov\text{-}snag)} = \frac{3[(80)(1.2)+(35)(3.3)+(5)(2.1)]-[(120)(6.6)]}{\{[3(80^2+35^2+5^2)-(120^2)]\,[3(1.2^2+3.3^2+2.1^2)-(6.6^2)]\}1/2} = -0.528$$

Correlation (r) Matrix

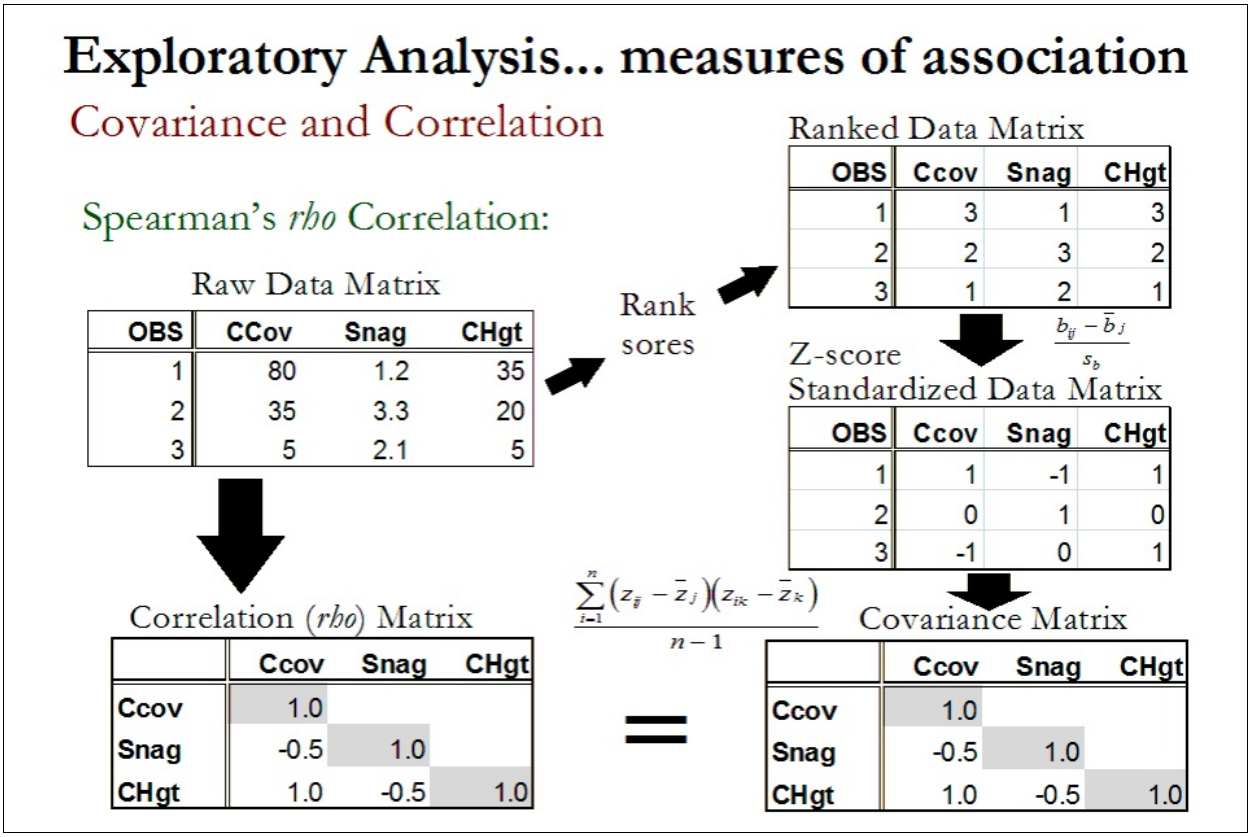
| | CCov | Snag | CHgt |
|---|---|---|---|
| **CCov** | 1.000 | | |
| **Snag** | -0.528 | 1.000 | |
| **CHgt** | 0.993 | -0.427 | 1.000 |

Diagonals = internal association
Off-diagonals = correlations

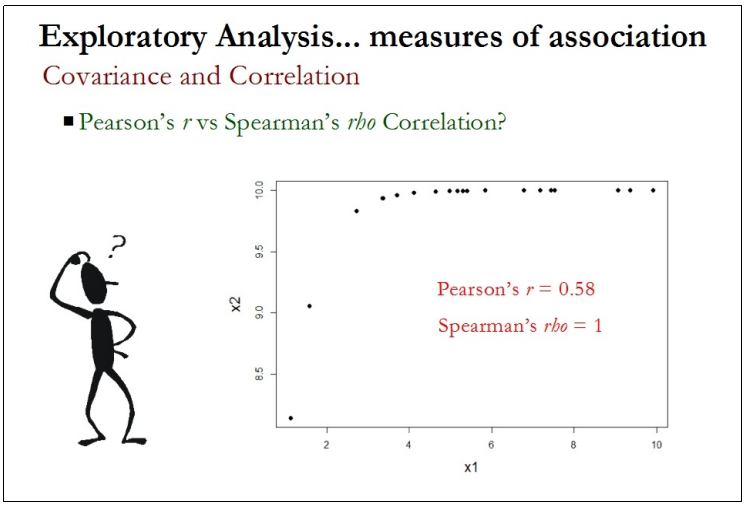
The calculation of Pearson's *r* correlation coefficient is shown here for the same data matrix. The formula for calculating Pearson's *r* is rather cumbersome and not intuitive, but is shown here for completeness. Note, the denominator of the equation a scaling factor that scales the result to range between -1 and 1, where a -1 is a perfect inverse (or negative) correlation and a +1 is a perfect (positive) correlation, and a 0 mean no correlation. A zero correlation indicates that the two variables are statistically independent; i.e., they do not covary. Like the covariance matrix, the correlation matrix is a square symmetric matrix with the correlation of a variable with itself (always a perfect positive correlation, or 1) along the diagonals and the pairwise correlation coefficients in the off-diagonal positions. Again, because the matrix is symmetric, only the lower triangle is shown.




It is important to understand the relationship between covariance and Pearson's correlation coefficient. The correlation between two variables is equal to their covariance computed on $z$-score standardized variables. If we first standardized each variable using a $z$-score standardization, which involves centering each variable on 0 (i.e., shifting the mean to zero) and scaling the spread of each distribution to unit variance (i.e., variance and standard deviation equal to 1), and then compute the regular covariance between the variables, we get the same result as computing the correlation on the raw data. Thus, correlation is simply standardized covariance. Whereas the covariance is unbounded and entirely depends on the scale of the variable, the correlation is always bounded by -1 and 1. Consequently, the correlation coefficient has a very straightforward and intuitive interpretation whereas the covariance does not.




In a similar manner, Spearman's *rho* correlation is simply the Pearson's correlation between rank-standardized variables. First, we standardized each variable using a rank standardization, which involves converting each data value to its rank.. Then, we standardize the ranked data using a z-score standardization and compute the regular covariance as before. Whereas Pearson's correlation coefficient does a good job of describing the linear association between two continuous variables, Spearman's correlation coefficient is much less sensitive to departures from linear association and thus is better for describing the monotonic relationship between two variables. By monotonic, I mean an always increasing or always decreasing relationship. Thus, relationships that are consistently positive or, conversely, consistently negative, but not necessarily perfectly linear, have a perfect Spearman's correlation of 1 (or conversely -1), as in the figure shown here. Thus, the choice between Pearson's *r* and Spearman's *rho* depends on whether we are interested in linear or monotonic associations.



# Exploratory Analysis… measures of association

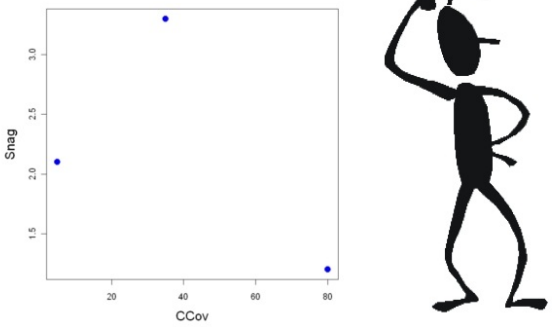
## Covariance and Correlation

Pearson's *r* Correlation

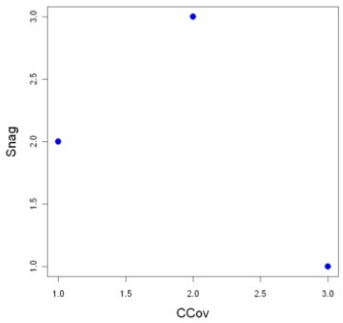
| | Ccov | Snag | CHgt |
|---|---|---|---|
| **Ccov** | 1.00 | | |
| **Snag** | -0.53 | 1.00 | |
| **CHgt** | 0.99 | -0.43 | 1.00 |

Spearman's *rho* Correlation

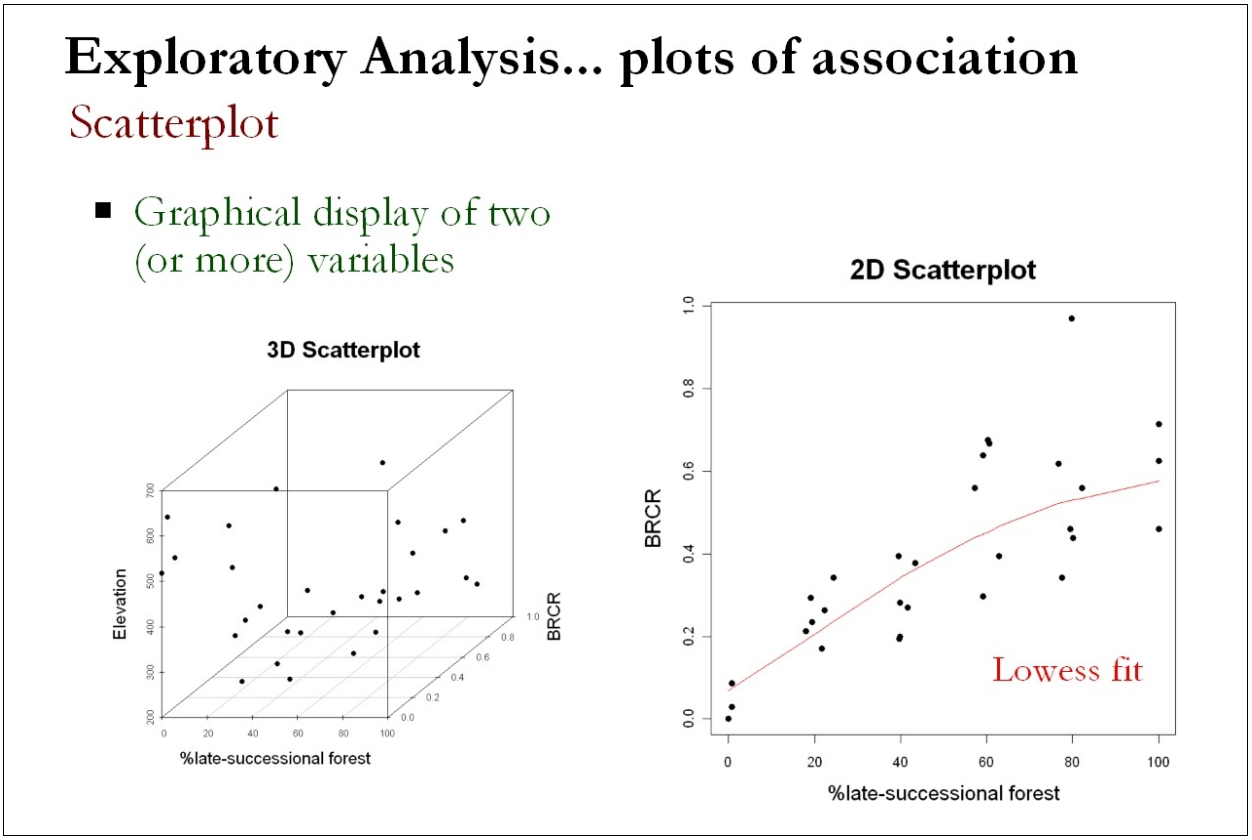
| | Ccov | Snag | CHgt |
|---|---|---|---|
| **Ccov** | 1.00 | | |
| **Snag** | -0.50 | 1.00 | |
| **CHgt** | 1.00 | -0.50 | 1.00 |






As shown here, in this particular data set, Pearson's *r* and Spearman's *rho* are quite similar and depict a negative association between CCov and Snag – as CCov increases Snag generally decreases, but the association is far from perfect, leading to -0.53 and -0.50 correlation coefficients, respectively. Notice in the scatterplots how the rank-transformed data associated with the Spearman's correlation has a different scale for the x and y axes; specifically, the points are positioned at their ranks: 1, 2, or 3, instead of their raw scores as in the scatterplot on the left associated with the Pearson's correlation.




## 5. Plots of association

*5.1 Scatterplot*

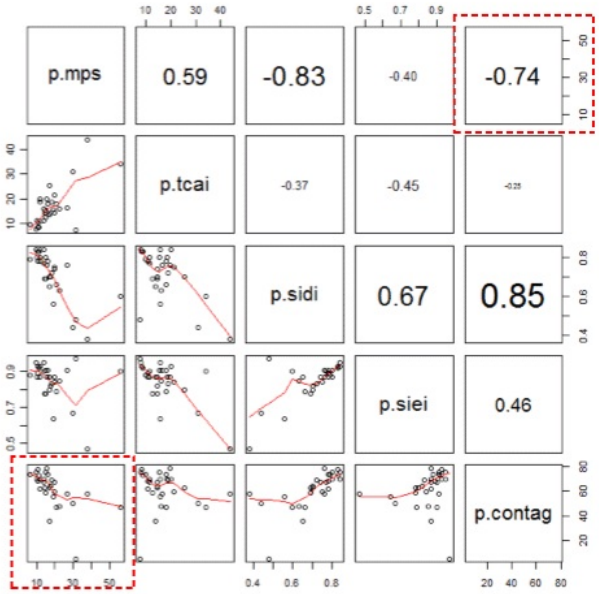
In cases involving continuous dependent and independent variables, it can be very useful to examine scatterplots between pairs of dependent and independent variables prior to constructing and anlayzing a statistical model. At the risk of data dredging, the graphical relationship display in the scatterplot can provide an indication of the strength and nature of the dependent relationship and thereby guide the selection of the appropriate statistical model to follow. The 2-dimensional scatterplot shown here is a graphical depiction of the relationship between a single independent variable (%late-successional forest) and a single dependent variable (BRCR, representing the relative abundance of brown creepers) for 30 landscapes in the Oregon Coast Range. Also shown is a robust locally weighted regression (lowess) line that depicts the general pattern in the data without being overly constrained by a specific statistical model of the relationship. Thus, the line can wiggle around as much as needed to reflect the general patterns in the data. As such, it can be a useful guide as to the shape of the underlying relationship and the form of the statistical model to be pursued later. Note, scatterplots can be equally useful for assessing relationships between independent variables, since many statistical methods assume that the independent variables themselves are truly independent of each or at least not strongly dependent. Scatterplots can be extended to a third dimension, as shown here on the left, in order to depict the relationship among three variables.

# Exploratory Analysis... plots of association

## Scatterplot Matrix

- Matrix of scatterplots for every combination of variables (and the corresponding correlation coefficients)

| | p.mps | p.tcai | p.sidi | p.siei | p.contag |
|---|---|---|---|---|---|
| 1 | 17.19 | 25.59 | 0.70 | 0.80 | 62.97 |
| 2 | 22.54 | 16.01 | 0.63 | 0.85 | 47.32 |
| 3 | 10.04 | 11.41 | 0.80 | 0.87 | 69.69 |
| 4 | 17.93 | 18.66 | 0.72 | 0.83 | 69.42 |
| 5 | 11.80 | 20.05 | 0.84 | 0.93 | 70.15 |
| 6 | 11.91 | 18.54 | 0.80 | 0.87 | 69.74 |
| 7 | 26.85 | 16.29 | 0.76 | 0.91 | 58.12 |
| 8 | 19.53 | 21.32 | 0.75 | 0.84 | 59.29 |
| 9 | 37.65 | 43.83 | 0.38 | 0.47 | 57.88 |
| 10 | 29.88 | 31.07 | 0.44 | 0.67 | 50.12 |




*5.2 Scatterplot matrix*

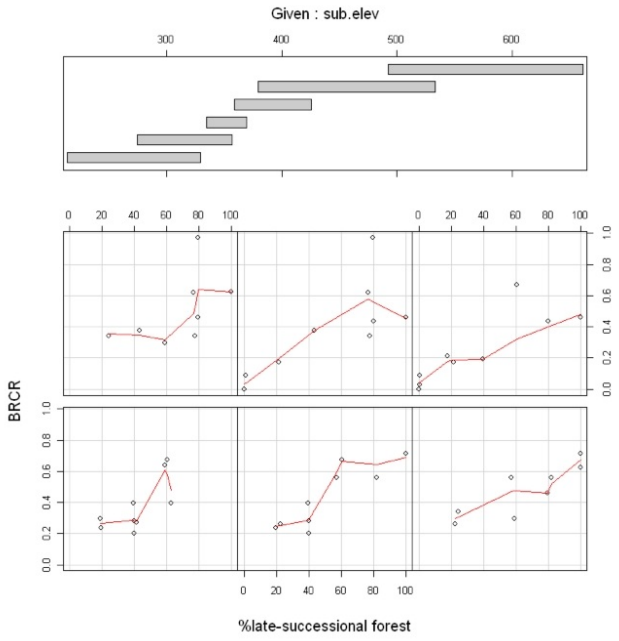
In cases involving many variables, it can also be quite useful (and efficient) to produce bi-variate (2D) scatterplots for all pairs of variables. A scatterplot matrix does just this. As shown in this example, the scatterplot matrix depicts variables along the diagonal, pairwise scatterplots in the lower triangle, and the corresponding correlation coefficient (of your choice, e.g., Pearson's *r* or Spearman's *rho*) in the upper triangle. Notice here that we also added the Lowess regression lines to each of the scatterplots to enhance the interpretation.

# Exploratory Analysis... plots of association

## Coplot

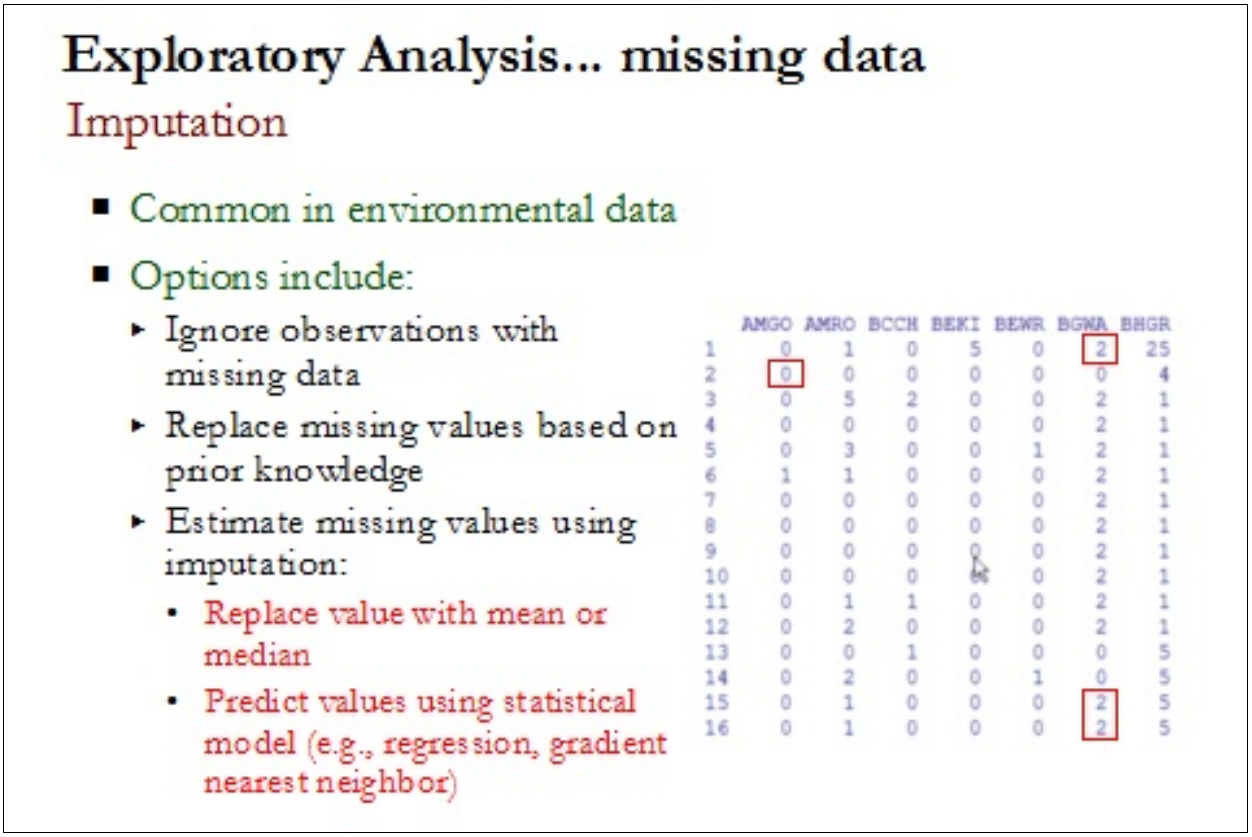
- Scatterplots of two variables conditioned on a third; i.e., relationship between $x$ and $y$, given $z$

Read from lower left to upper right



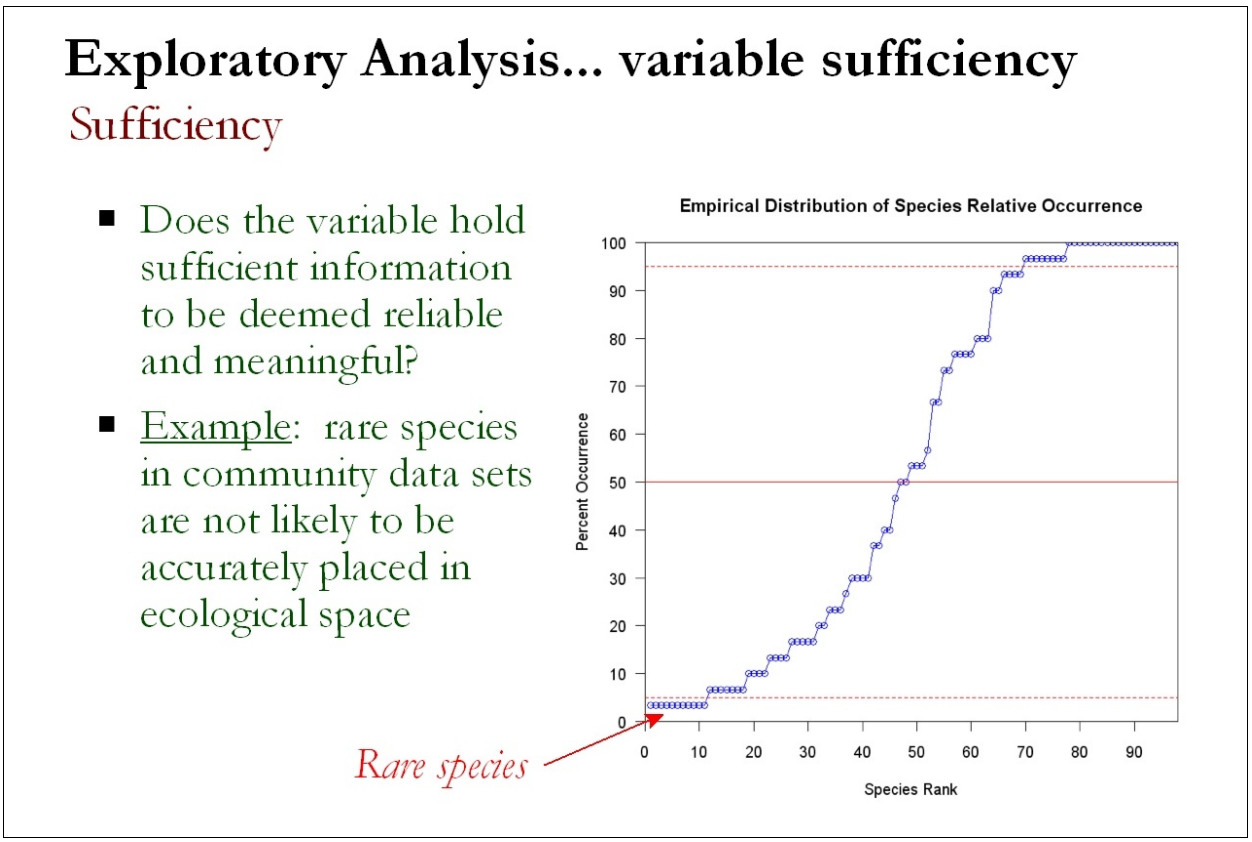

*5.3 Coplot*

In cases involving the relationship between a dependent and independent variable, the relationship may be obscured by the effects of other variables. In such cases it may be useful to examine a scatterplot of x and y, but conditioned on a third variable, say $z$. A coplot does just this. In the example shown here, the panels in the coplot are ordered from lower left to upper right, associated with the values of the conditioning variable in the upper panel, read left to right. So, the lower left scatterplot is for points corresponding to the leftmost bar in the upper panel. In this case, the lower left panel shows the scatterplot of %late-successional forest versus brown creeper abundance for the nine landscapes at the lowest elevations. The next plot to the right shows the same thing but for nine landscapes centered on slightly higher elevations. The top-right scatterplot shows the landscapes at the highest elevations. In this manner, the scatterplot of %late-successional forest and brown creeper abundance is "conditioned" on elevation.




## 6. Missing data

Despite the best laid plans, environmental data often contain missing data. Missing data can arise for all sorts of reasons, the problem is what to do with it? There are lots of options for dealing with missing data, ranging from simple to complex. Perhaps the easiest solution is to ignore or delete observations with any missing data. This is a luxury we often cannot afford since we may have a small sample size to begin with. Another option is to replace the missing values with values based on expert prior knowledge. This of course is risky business and should not be done unless under very special circumstances. A final solution is to estimate the missing values using methods of imputation. The simplest of these, and therefore the most commonly used, is to replace the missing value with the mean or median of the variable. The purpose behind this imputation method is to replace the missing value with a value that will no exert any influence on the analysis. There are much more complex methods of imputation, including for example using a statistical model to predict the missing values based on the other variables in the data set. This procedure comes at the cost of using the same data to predict the missing values as we intend to use in our final statistical model. One solution of course is to use a separate set of variables for the imputation than we intend to use in the final model. Regardless of the method employed, we have to be suspicious of any data set in which a large number of missing values have been replaced.

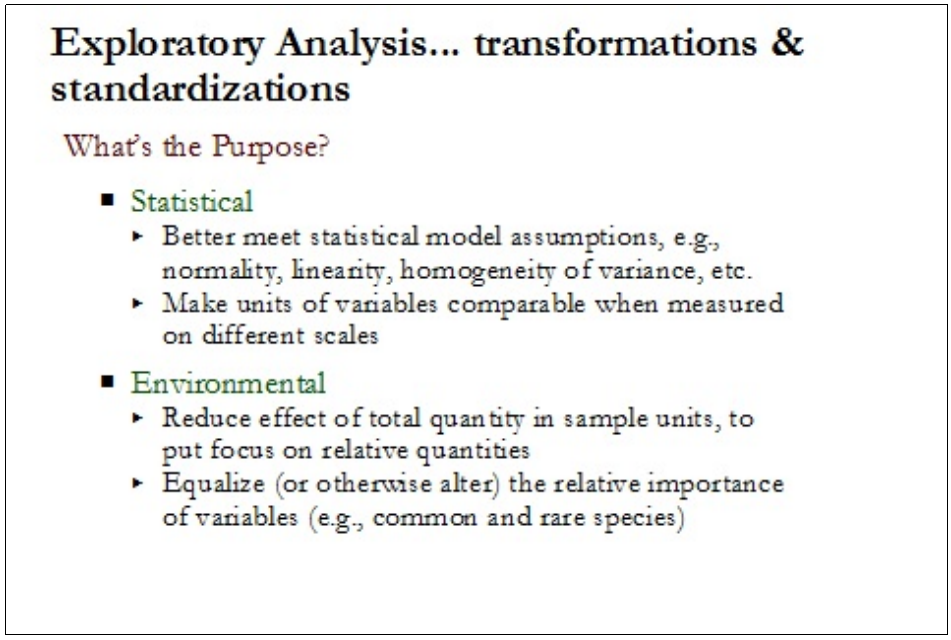



## 7. Variable Sufficiency

An often overlooked but important step prior to statistical modeling is to screen the data for insufficient variables (i.e., those that were sampled insufficiently to reliably characterize their environmental pattern). For example, in community data sets rare species with very few records are not likely to be accurately placed in ecological space. We must decide at what level of frequency of occurrence we want to accept the 'message' and eliminate species below this level.

In the example shown here, 98 breeding birds species were detected across 30 landscapes in the Oregon Coast Range. The x-axis lists species in their rank order of percent occurrence across the 30 landscapes, so the first point is for the species with the lowest percent occurrence, here corresponding to 3.3% (or 1/30 landscapes). The plot reveals that there are 11 species with <5% occurrence. It seems unlikely that we will be able to model these species patterns of occurrence in relationship to the other species or the habitat variables (not shown) based on a single occurrence. The information in this data set is insufficient to reliable model these species and therefore they should be dropped before further analysis. Unfortunately, there is no objective threshold for determining when there is sufficient information on a variable, so we must rely on intuition. In community data sets it is quite common to drop rare species occurring on fewer than say four plots.

## 8. Data transformations and standardizations



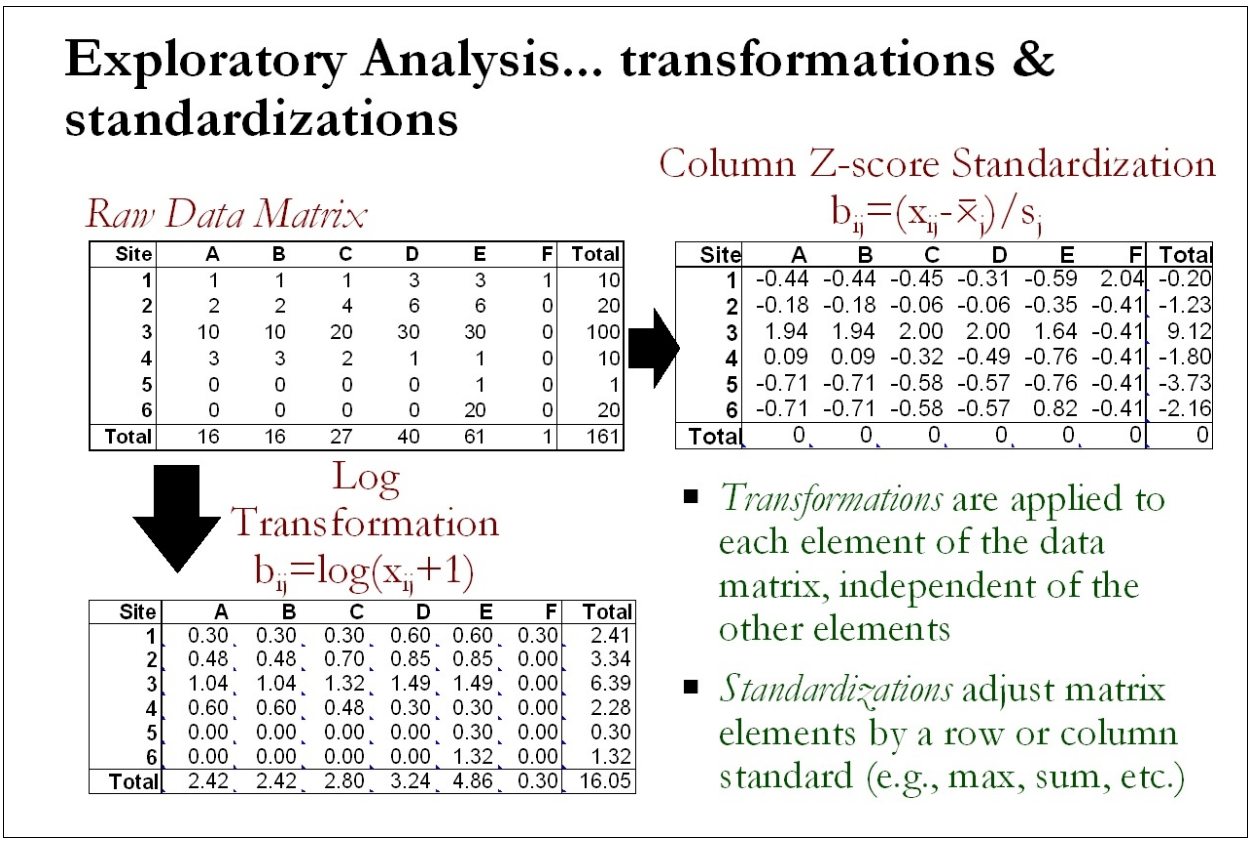

Once we have thoroughly screened our data, we may find it useful or necessary to transform and/or standardize the data. There are both statistical and environmental reasons for considering adjustment of the data:

Statistical reasons:

- Better meet statistical model assumptions, e.g., normality, linearity, homogeneity of variance, etc.. Typically, data adjustments for this purpose are made after the initial modeling has revealed a problem and it is believed that an adjustment of the data might solve the problem.
- Make units of variables comparable when measured on different scales. This is a common situation in environmental data sets where the variables are often measured on wildly different scales (e.g., pH, percent cover, mass, etc.). However, the need to adjust data to account for these different scales entirely depends on the statistical model and method used.

Environmental reasons:

- Reduce the effect of total quantity in sample units, to put the focus on relative quantities. This is a common reason in community data sets involving sites by species data where the sites may vary dramatically in total species abundance but the pattern that is of interest is the relative abundance of the constituent species.
- Equalize (or otherwise alter) the relative importance of variables (e.g., common and rare species). This too is a common reason in community data sets involving sites by species data where the species may vary dramatically in their total abundance across sites but the pattern that is of interest is their relative abundance profiles across sites.




| Site | A | B | C | D | E | F | Total |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 3 | 3 | 1 | 10 |
| 2 | 2 | 2 | 4 | 6 | 6 | 0 | 20 |
| 3 | 10 | 10 | 20 | 30 | 30 | 0 | 100 |
| 4 | 3 | 3 | 2 | 1 | 1 | 0 | 10 |
| 5 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| 6 | 0 | 0 | 0 | 0 | 20 | 0 | 20 |
| Total | 16 | 16 | 27 | 40 | 61 | 1 | 161 |

| Site | A | B | C | D | E | F | Total |
|---|---|---|---|---|---|---|---|
| 1 | -0.44 | -0.44 | -0.45 | -0.31 | -0.59 | 2.04 | -0.20 |
| 2 | -0.18 | -0.18 | -0.06 | -0.06 | -0.35 | -0.41 | -1.23 |
| 3 | 1.94 | 1.94 | 2.00 | 2.00 | 1.64 | -0.41 | 9.12 |
| 4 | 0.09 | 0.09 | -0.32 | -0.49 | -0.76 | -0.41 | -1.80 |
| 5 | -0.71 | -0.71 | -0.58 | -0.57 | -0.76 | -0.41 | -3.73 |
| 6 | -0.71 | -0.71 | -0.58 | -0.57 | 0.82 | -0.41 | -2.16 |
| Total | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

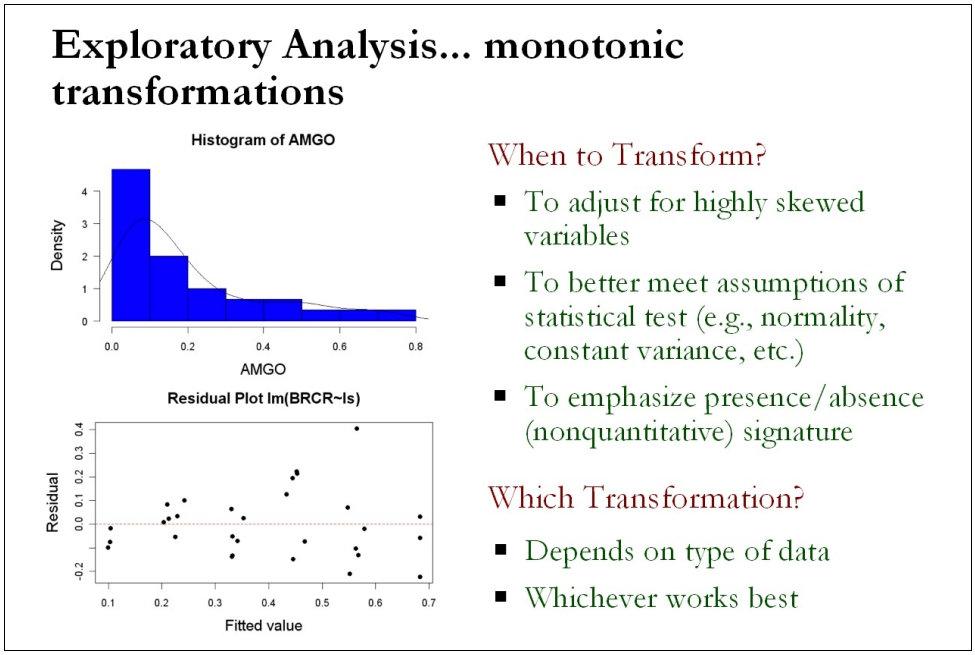
| Site | A | B | C | D | E | F | Total |
|---|---|---|---|---|---|---|---|
| 1 | 0.30 | 0.30 | 0.30 | 0.60 | 0.60 | 0.30 | 2.41 |
| 2 | 0.48 | 0.48 | 0.70 | 0.85 | 0.85 | 0.00 | 3.34 |
| 3 | 1.04 | 1.04 | 1.32 | 1.49 | 1.49 | 0.00 | 6.39 |
| 4 | 0.60 | 0.60 | 0.48 | 0.30 | 0.30 | 0.00 | 2.28 |
| 5 | 0.00 | 0.00 | 0.00 | 0.00 | 0.30 | 0.00 | 0.30 |
| 6 | 0.00 | 0.00 | 0.00 | 0.00 | 1.32 | 0.00 | 1.32 |
| Total | 2.42 | 2.42 | 2.80 | 3.24 | 4.86 | 0.30 | 16.05 |

It is important to distinguish between a "transformation" and a "standardization" as these terms are often confused and used in potentially misleading wasys.

A data *transformation* involves applying a mathematical function separately to each data value (i.e., a single cell in the data frame). Each cell is transformed in isolation and independently of any other cell or any other information in the data set. For example, a log transformation involves returning the logarithm of the cell value, which depends only on the cell value itself.

A data *standardization* (sometimes also referred to as "relativization") involves adjusting a data value relative to a specified standard derived from the corresponding row (sample) and/or column (variable) of the data frame. For example, dividing each cell value by the total or sum of that variable across all samples is a standardization because the standard is derived from information outside of the focal cell.

## *8.1. Monotonic Transformations*




The transformations commonly used with environmental data, including all of those considered below are monotonic; that is, the transformation does not change the rank ordering of values.

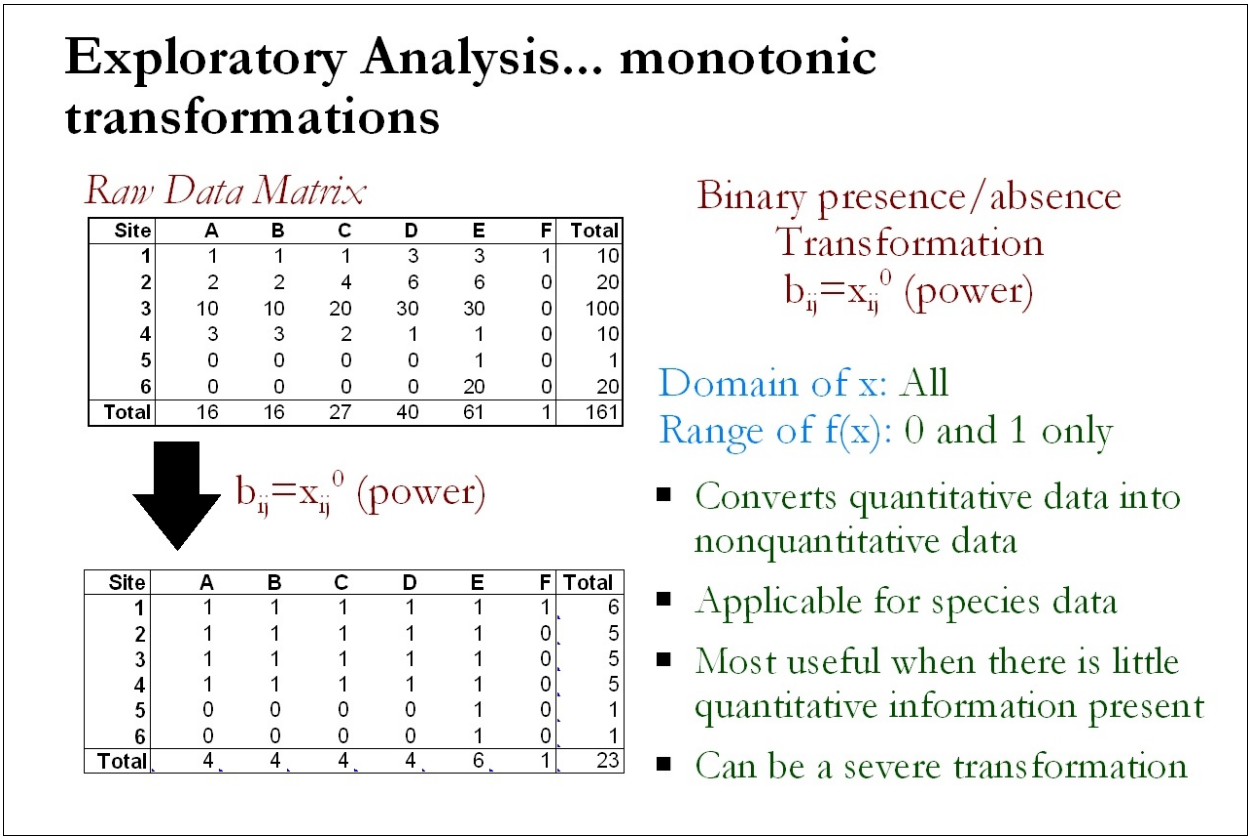
When to transform?

The most difficult aspect of data adjustment is knowing when and when not to transform (and/or standardize) the data. Often times environmental data are highly skewed and/or range over several orders of magnitude, and as such can benefit from a transformation, such as the log or square-root transformation, that compress large values. For community data sets involving species abundances, it is sometimes useful or more meaningful to transform the data to binary (presence/absence) data. Here are some general rules for when to transform:

- To adjust for highly skewed variables. Sometimes it is necessary to make distributions more symmetrical to better the assumptions of particular statistical tests. Environmental data often contain positively skewed distributions which can in some cases be problematic for statistical models. Some transformations act to pull the tail of the distribution in and in so doing reduce skew.
- To better meet assumptions of statistical test (e.g., normality, constant variance, etc.). Parametric statistical models come with sometimes onerous assumptions regarding the distribution of the data and transformation can sometimes help us better meet those assumptions.
- To emphasize presence/absence (nonquantitative) signature. In some environmental data sets, especially community data sets, the dominant pattern of interest may be the presence/absence of an attribute (e.g., species present or absent) rather than the quantitative data collected. Transformations can convert the data from quantitative to binary present/absent.

Which transformation?

Another difficult decision is which transformation to use to achieve the stated purpose. Ultimately the decision depends on the type of data involved but in many cases it is simply a matter of determining post-hoc which transformation works best. Some general rules of thumb are given below.




*Binary transformation*

Any quantitative data can be transformed to binary present/absent data by taking the value raised to the zero power. Hence, the binary transformation is actually a special case of the power transformation (see below) when the power is zero.

The acceptable domain of x (i.e., acceptable values of the raw data) is anything and the transformation returns a 0 or 1. The binary transformation converts quantitative data into nonquantitative data; is especially applicable for species data; is most useful when there is little quantitative information present in the variable; and can be a severe transformation since all the quantitative information is removed from the variable.

## Exploratory Analysis... monotonic transformations

*Raw Data Matrix*

| Site | A | B | C | D | E | F | Total |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 3 | 3 | 1 | 10 |
| 2 | 2 | 2 | 4 | 6 | 6 | 0 | 20 |
| 3 | 10 | 10 | 20 | 30 | 30 | 0 | 100 |
| 4 | 3 | 3 | 2 | 1 | 1 | 0 | 10 |
| 5 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| 6 | 0 | 0 | 0 | 0 | 20 | 0 | 20 |
| Total | 16 | 16 | 27 | 40 | 61 | 1 | 161 |

$b_{ij}=\log(x_{ij}+1)$ ?

| Site | A | B | C | D | E | F | Total |
|---|---|---|---|---|---|---|---|
| 1 | 0.30 | 0.30 | 0.30 | 0.60 | 0.60 | 0.30 | 2.41 |
| 2 | 0.48 | 0.48 | 0.70 | 0.85 | 0.85 | 0.00 | 3.34 |
| 3 | 1.04 | 1.04 | 1.32 | 1.49 | 1.49 | 0.00 | 6.39 |
| 4 | 0.60 | 0.60 | 0.48 | 0.30 | 0.30 | 0.00 | 2.28 |
| 5 | 0.00 | 0.00 | 0.00 | 0.00 | 0.30 | 0.00 | 0.30 |
| 6 | 0.00 | 0.00 | 0.00 | 0.00 | 1.32 | 0.00 | 1.32 |
| Total | 2.42 | 2.42 | 2.80 | 3.24 | 4.86 | 0.30 | 16.05 |

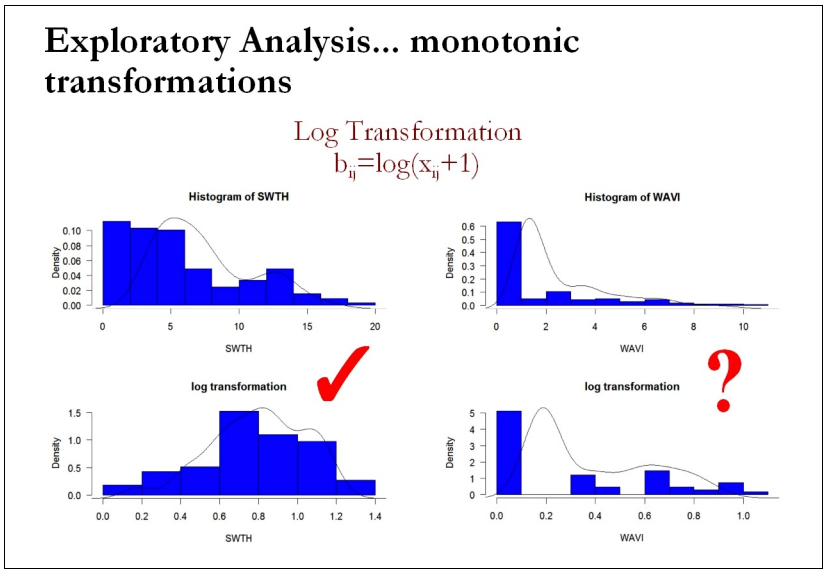
Log Transformation
$b_{ij}=\log(x_{ij}+1)$

Domain of x: >0
Range of f(x): All

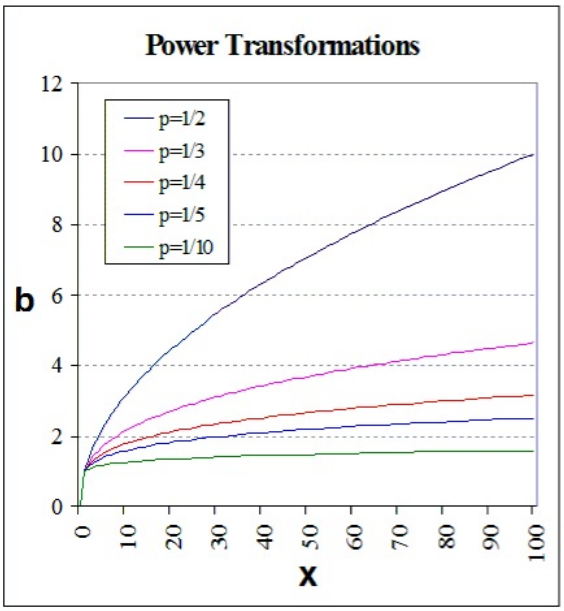
- Compresses high values and spreads low values by expressing values as orders of magnitude
- Useful when high degree of variation; ratio of largest to smallest >10; highly positively skewed data

*Log transformation*

The log transformation is very common in environmental data. The acceptable domain of x is non-zero positive values (note, the log of zero is undefined) and the transformation returns any real number (positive or negative). The log transformation compresses high values and spreads low values by expressing values as orders of magnitude; is very useful when there is a high degree of variation among the values (e.g., ratio of largest to smallest >10); and is often used to adjust highly positively skewed data.



## Exploratory Analysis... monotonic transformations




Power Family Transformation

$b_{ij}=x_{ij}^{p}$

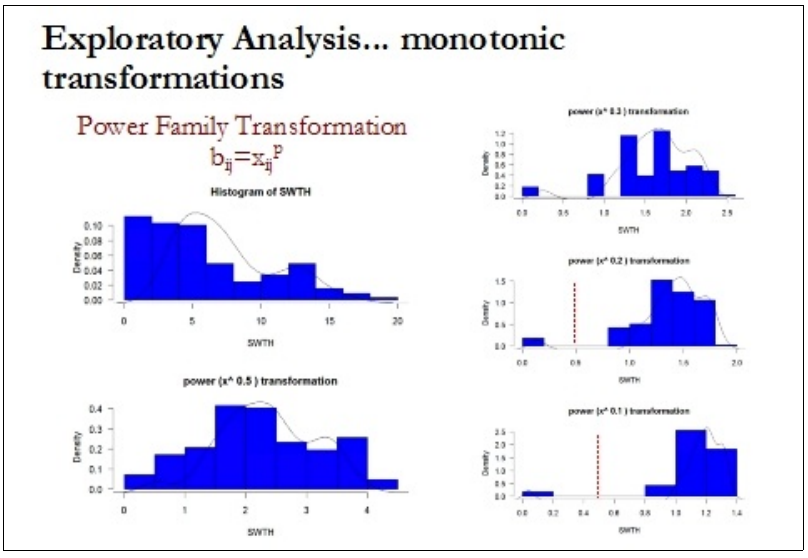
Domain of x: ≥0

Range of f(x): ≥0

- Different exponents change the effect of the transformation; the smaller the exponent, the more compression applied to high values
- Flexible transformation useful for a wide variety of data

*Power transformation*

The power transformation is a versatile transformation that involves raising the value to any specified power, usually less than 1. The domain of x is ≥0 and the transformation, regardless of power, returns a value in the same range. The power transformation has a varying effect depending on the power used; different exponents change the effect of the transformation; the smaller the exponent, the more compression applied to high values. Consequently, the power transformation is a flexible transformation useful for a wide variety of environmental data. Note, the square-root transformation is simply a special case of the power transformation when the exponent is equal to 0.5.



## Exploratory Analysis... monotonic transformations

*Raw Data Matrix*

| Site | A | B | C | D | E | F | Total |
|---|---|---|---|---|---|---|---|
| 1 | 0.06 | 0.06 | 0.04 | 0.08 | 0.05 | 1.00 | 1.29 |
| 2 | 0.13 | 0.13 | 0.15 | 0.15 | 0.10 | 0.00 | 0.65 |
| 3 | 0.63 | 0.63 | 0.74 | 0.75 | 0.49 | 0.00 | 3.23 |
| 4 | 0.19 | 0.19 | 0.07 | 0.03 | 0.02 | 0.00 | 0.49 |
| 5 | 0.00 | 0.00 | 0.00 | 0.00 | 0.02 | 0.00 | 0.02 |
| 6 | 0.00 | 0.00 | 0.00 | 0.00 | 0.33 | 0.00 | 0.33 |
| Total | 1 | 1 | 1 | 1 | 1 | 1 | 6 |

$b_{ij}=\log(x_{ij}/(1-x_{ij}))$

| Site | A | B | C | D | E | F | Total |
|---|---|---|---|---|---|---|---|
| 1 | -2.75 | -2.75 | -3.18 | -2.44 | -2.94 | Inf | -14.1 |
| 2 | -1.90 | -1.9 | -1.73 | -1.73 | -2.2 | -Inf | -9.47 |
| 3 | 0.53 | 0.53 | 1.05 | 1.05 | -0.04 | -Inf | 3.12 |
| 4 | -1.52 | -1.52 | -2.59 | -3.48 | -3.89 | -Inf | -13 |
| 5 | -Inf | -Inf | -Inf | -Inf | -3.89 | -Inf | -3.89 |
| 6 | -Inf | -Inf | -Inf | -Inf | -0.75 | -Inf | -0.75 |
| Total | -5.64 | -5.64 | -6.45 | -6.61 | -13.7 | n/a | -38.1 |

Logit Transformation
$b_{ij}=\log(x_{ij}/(1-x_{ij}))$

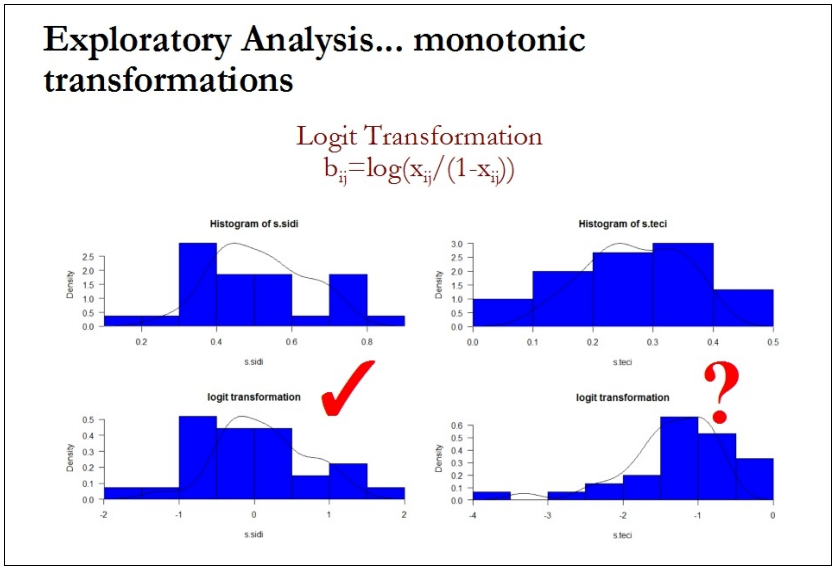
Domain of x: 0-1
Range of f(x): -∞ - ∞

- Spreads end of the scale while compressing the middle for proportion data
- Useful for proportion data to create unbounded distribution

*Logit transformation*

For data expressed as a proportion (i.e., range 0-1), the logit transformation is often recommended by statisticians. The acceptable domain of x is 0-1 and the transformation returns a value in an unbounded range (-∞ to ∞).

The effect of the transformation is to spread the end of the scale while compressing the middle, which can be quite useful for proportion data when the desire is convert a bounded distribution to an unbounded one, which can affect the choice of the appropriate probability distribution in the parametric statistic model.



## Exploratory Analysis... monotonic transformations

*Raw Data Matrix*

| Site | A | B | C | D | E | F | Total |
|---|---|---|---|---|---|---|---|
| 1 | 0.06 | 0.06 | 0.04 | 0.08 | 0.05 | 1.00 | 1.29 |
| 2 | 0.13 | 0.13 | 0.15 | 0.15 | 0.10 | 0.00 | 0.65 |
| 3 | 0.63 | 0.63 | 0.74 | 0.75 | 0.49 | 0.00 | 3.23 |
| 4 | 0.19 | 0.19 | 0.07 | 0.03 | 0.02 | 0.00 | 0.49 |
| 5 | 0.00 | 0.00 | 0.00 | 0.00 | 0.02 | 0.00 | 0.02 |
| 6 | 0.00 | 0.00 | 0.00 | 0.00 | 0.33 | 0.00 | 0.33 |
| Total | 1 | 1 | 1 | 1 | 1 | 1 | 6 |

$b_{ij}=(2/\pi)*\sin^{-1}(x_{ij}^{1/2})$

| Site | A | B | C | D | E | F | Total |
|---|---|---|---|---|---|---|---|
| 1 | 0.16 | 0.16 | 0.12 | 0.18 | 0.14 | 1.00 | 1.76 |
| 2 | 0.23 | 0.23 | 0.25 | 0.25 | 0.20 | 0.00 | 1.17 |
| 3 | 0.58 | 0.58 | 0.66 | 0.67 | 0.49 | 0.00 | 2.98 |
| 4 | 0.29 | 0.29 | 0.18 | 0.10 | 0.08 | 0.00 | 0.93 |
| 5 | 0.00 | 0.00 | 0.00 | 0.00 | 0.08 | 0.00 | 0.08 |
| 6 | 0.00 | 0.00 | 0.00 | 0.00 | 0.39 | 0.00 | 0.39 |
| Total | 1.256 | 1.256 | 1.21 | 1.198 | 1.392 | 1 | 7.3125 |

Arcsin Square Root Transformation

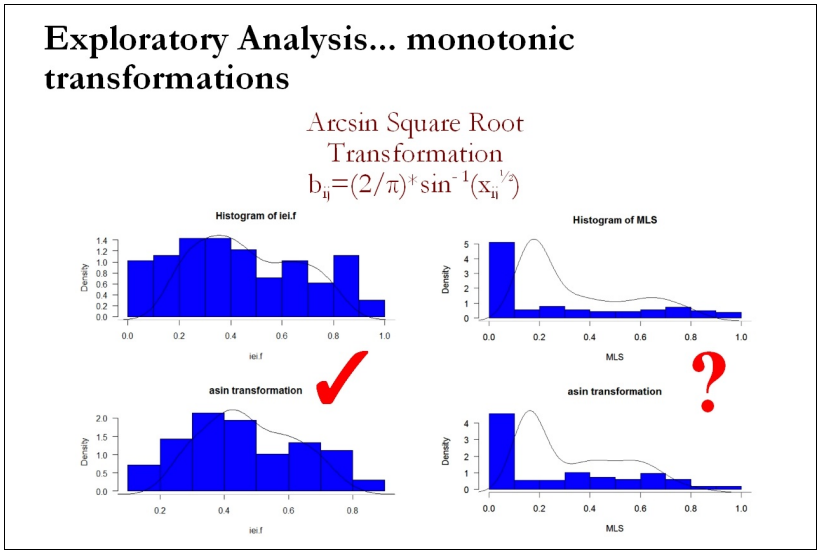
$b_{ij}=(2/\pi)*\sin^{-1}(x_{ij}^{1/2})$

Domain of x: 0-1
Range of f(x): 0-1

- Spreads end of the scale while compressing the middle for proportion data
- Useful for proportion data with positive skew (can use arcsine transformation for negative skew)

*Arcsine square-root transformation*

For data expressed as a proportion (i.e., ranges 0-1), the arcsine square-root transformation is also often recommended by statisticians. The acceptable domain of x is 0-1 and the transformation returns a value in the same range (0-1). Like the logit transformation, the effect of the transformation is to spread the end of the scale while compressing the middle, only the arcsine square-root transformation maintains the original 0-1 range of the data. This transformation can be useful for proportion data with positive skew. Note, the arcsine transformation (minus the square root) can be used for negative skew.



# Exploratory Analysis... monotonic transformations

*Raw Data Matrix*

| Site | A | B | C | D | E | F | Total |
|---|---|---|---|---|---|---|---|
| 1 | 0 | 180 | 0 | 55 | 0 | 90 | 325 |
| 2 | 10 | 45 | 0 | 25 | 0 | 110 | 190 |
| 3 | 33 | 45 | 120 | 5 | 0 | 75 | 278 |
| 4 | 45 | 0 | 270 | 0 | 355 | 80 | 750 |
| 5 | 90 | 280 | 225 | 10 | 340 | 45 | 990 |
| 6 | 355 | 35 | 330 | 0 | 15 | 95 | 830 |
| Total | 533 | 585 | 945 | 95 | 710 | 495 | 3363 |

$b_{ij}=(\cos(\text{rad}(0-x_{ij}))+1)/2$

| Site | A | B | C | D | E | F | Total |
|---|---|---|---|---|---|---|---|
| 1 | 1.00 | 0.00 | 1.00 | 0.79 | 1.00 | 0.50 | 4.287 |
| 2 | 0.99 | 0.85 | 1.00 | 0.95 | 1.00 | 0.33 | 5.128 |
| 3 | 0.92 | 0.85 | 0.25 | 1.00 | 1.00 | 0.63 | 4.65 |
| 4 | 0.85 | 1.00 | 0.50 | 1.00 | 1.00 | 0.59 | 4.938 |
| 5 | 0.50 | 0.59 | 0.15 | 0.99 | 0.97 | 0.85 | 4.049 |
| 6 | 1.00 | 0.91 | 0.93 | 1.00 | 0.98 | 0.46 | 5.28 |
| Total | 5.263 | 4.204 | 3.829 | 5.73 | 5.951 | 3.355 | 28.33 |

Cosine Transformation

$b_{ij}=(\cos(\text{rad}(z-x_{ij}))+1)/2$

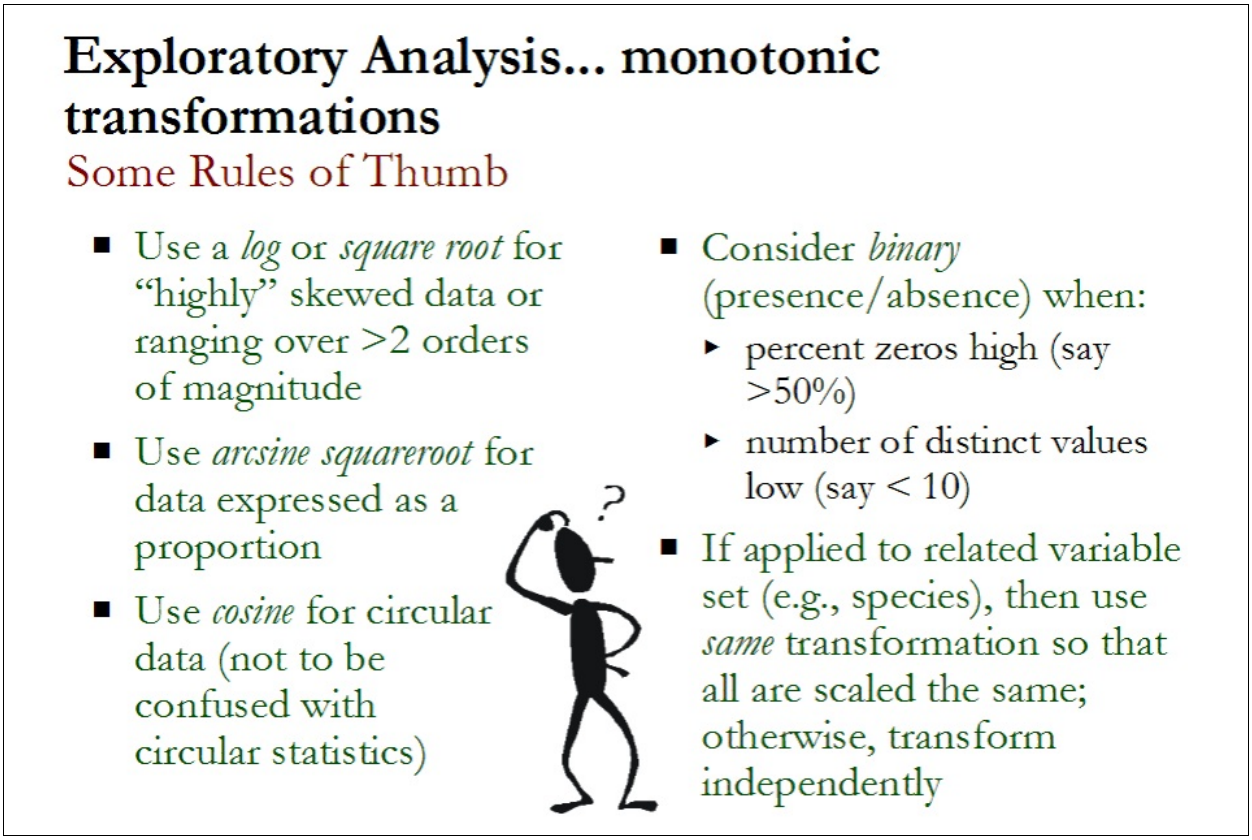
z=reference axis degrees

Domain of x: 0-360 degrees
Range of f(x): 0-1

- Converts circular measure (aspect) into linear gradient along specified reference axis
- Necessary for circular aspect data

*Cosine transformation*

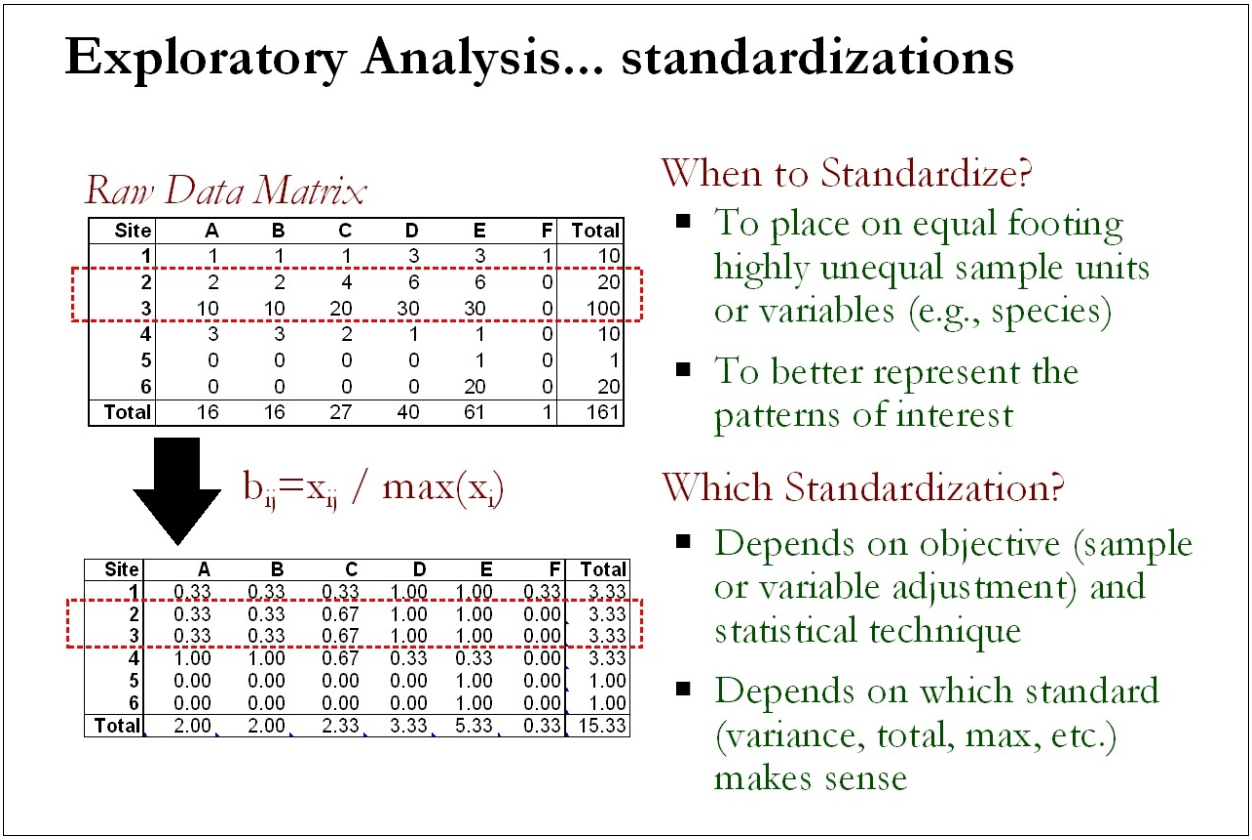
For circular data expressed in degrees, where the beginning and ending value of the numeric sequence (0 and 360) are equivalent, transformation is necessary prior to statistical modeling. The exception is with special statistical methods design specifically for such data. However, to use most conventional methods, the circular data requires transformation. The cosine transformation is the most commonly used and this involves first converting decimal degrees to radians and then taking the cosine of the radians. The +1 and /2 in the equation are used to scale the result to 0-1. The acceptable domain of x is 0-360 degree and the transformation returns a value between 0-1. The cosine transformation converts decimal degrees into a linear gradients defined along a specified reference axis. The $z$ parameter in the formula is used to specify the reference axis in decimal degrees. For example, if $z$=0 the reference axis is oriented north-south and the transformation will return values approach 1 as the angles approach the north and 0 as the angles approach the south from either direction. This transformation is common used to convert slope aspect into a usable form.




*Some rules of thumb*

Here are some general rules of thumb for using transformations:

- Use a log or square root for "highly" skewed data or ranging over >2 orders of magnitude.
- Use arcsine squareroot for data expressed as a proportion.
- Use cosine for circular data.
- Consider binary (presence/absence) transformation when either the percent zeros high (say >50%) or the number of distinct values is low (say < 10)
- If applied to related variable set (e.g., species), then use same transformation so that all are scaled the same; otherwise, transform independently.




*8.2 Standardizations*

In many environmental data sets, especially community data sets involving species abundances, it is often quite useful to standardize (or relativize) the data before conducting subsequent analyses. Recall that data standardization involves adjusting a data value relative to a specified standard derived from the corresponding row (sample) or column (variable) of the data set. Keep in mind that standardizations can fundamentally alter the patterns in the data and can make the difference "between illusion and insight, fog and clarity" (McCune and Grace, 2002).

When to standardize?
Knowing when to standardize is exceedingly difficult; recall the general purposes stated previously:

- Make units of variables comparable when measured on different scales.
- Reduce the effect of total quantity in sample units, to put the focus on relative quantities.
- Equalize (or otherwise alter) the relative importance of variables (e.g., common and rare species).

Which standardization?
An even more difficult decision is which standardization to use to achieve the stated purpose. Ultimately the decision depends on the objective (e.g., sample or variable adjustment), the subsequent statistical method used, and which standard (as the basis for the adjustment) makes the most sense. Unfortunately, wisdom in this regard only comes through experience.

# Exploratory Analysis... standardizations

## Row or Column Standardization

*Raw Data Matrix*

| Site | A | B | C | D | E | F | Total |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 3 | 3 | 1 | 10 |
| 2 | 2 | 2 | 4 | 6 | 6 | 0 | 20 |
| 3 | 10 | 10 | 20 | 30 | 30 | 0 | 100 |
| 4 | 3 | 3 | 2 | 1 | 1 | 0 | 10 |
| 5 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| 6 | 0 | 0 | 0 | 0 | 20 | 0 | 20 |
| Total | 16 | 16 | 27 | 40 | 61 | 1 | 161 |

$b_{ij}=(x_{ij}-\bar{x}_i)/s_i$

| Site | A | B | C | D | E | F | Total |
|---|---|---|---|---|---|---|---|
| 1 | -0.71 | -0.71 | -0.71 | 1.41 | 1.41 | -0.71 | 0.00 |
| 2 | -0.60 | -0.60 | 0.30 | 1.21 | 1.21 | -1.51 | 0.00 |
| 3 | -0.60 | -0.60 | 0.30 | 1.21 | 1.21 | -1.51 | 0.00 |
| 4 | 1.21 | 1.21 | 0.30 | -0.60 | -0.60 | -1.51 | 0.00 |
| 5 | -0.45 | -0.45 | -0.45 | -0.45 | 2.24 | -0.45 | 0.00 |
| 6 | -0.45 | -0.45 | -0.45 | -0.45 | 2.24 | -0.45 | 0.00 |
| Total | -1.6 | -1.6 | -0.7 | 2.329 | 7.695 | -6.12 | 0 |

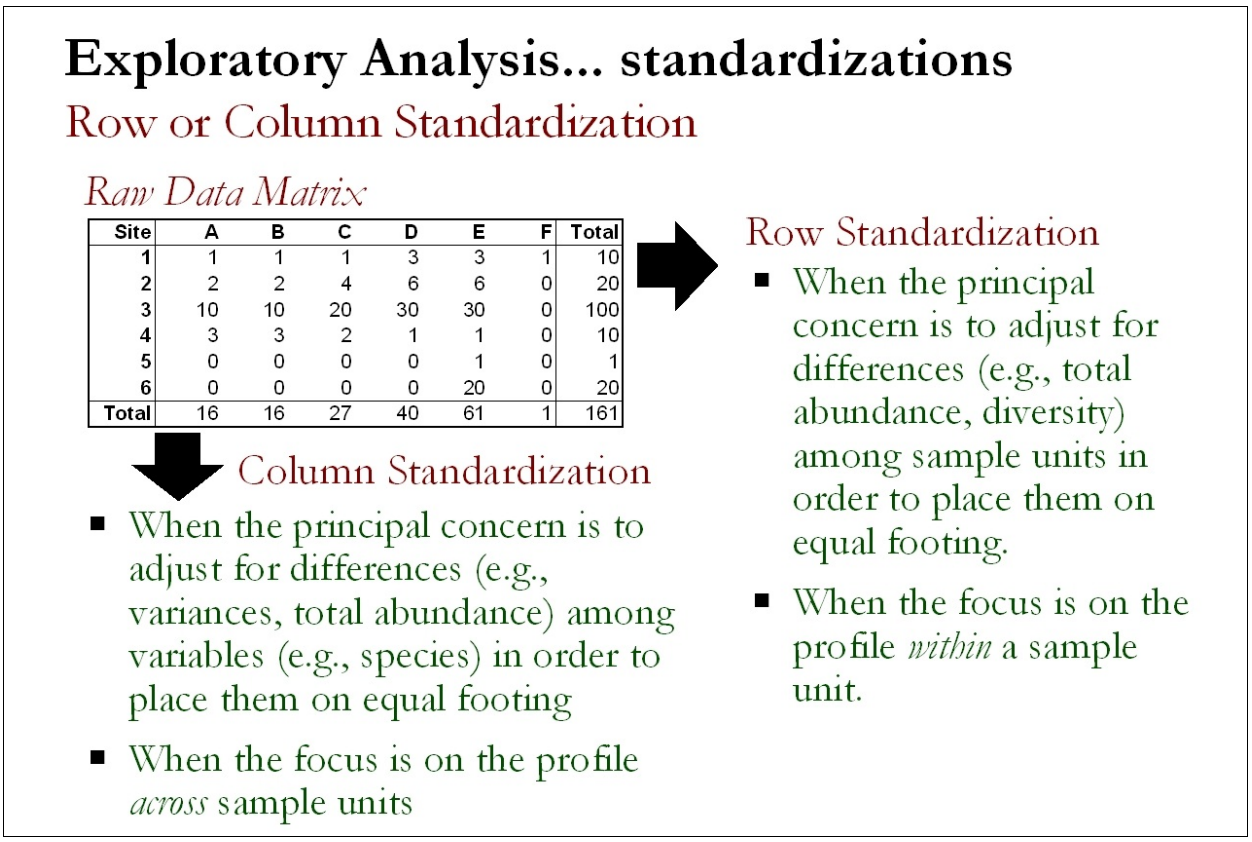
$b_{ij}=(x_{ij}-\bar{x}_j)/s_j$

| Site | A | B | C | D | E | F | Total |
|---|---|---|---|---|---|---|---|
| 1 | -0.48 | -0.48 | -0.50 | -0.34 | -0.65 | 2.24 | -0.22 |
| 2 | -0.19 | -0.19 | -0.07 | -0.06 | -0.38 | -0.45 | -1.35 |
| 3 | 2.13 | 2.13 | 2.19 | 2.19 | 1.80 | -0.45 | 10.00 |
| 4 | 0.10 | 0.10 | -0.35 | -0.53 | -0.83 | -0.45 | -1.97 |
| 5 | -0.77 | -0.77 | -0.64 | -0.63 | -0.83 | -0.45 | -4.09 |
| 6 | -0.77 | -0.77 | -0.64 | -0.63 | 0.89 | -0.45 | -2.36 |
| Total | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

- Standardizations adjust matrix elements by a row or column standard (e.g., max, sum, etc.)
- All standardizations can be applied to either rows or columns (or both)

It is important to remember that standardizations adjust data elements by a row or column standard (e.g., max, sum, etc.), in contrast to transformation which depend on no standard. In addition, all standardizations can be applied to either rows or columns (or both).

In the example shown here, the raw data matrix contains 6 sample plots (rows) and 6 species variables (columns). The standardization employed is the *z*-score standardization which involves centering the values on zero and scaling the spread to 1. This is accomplished by subtracting the mean from each value and dividing by the standard deviation. This *z*-score standardization can be applied to each column (lower left matrix) or each row (upper right matrix), with very different results.

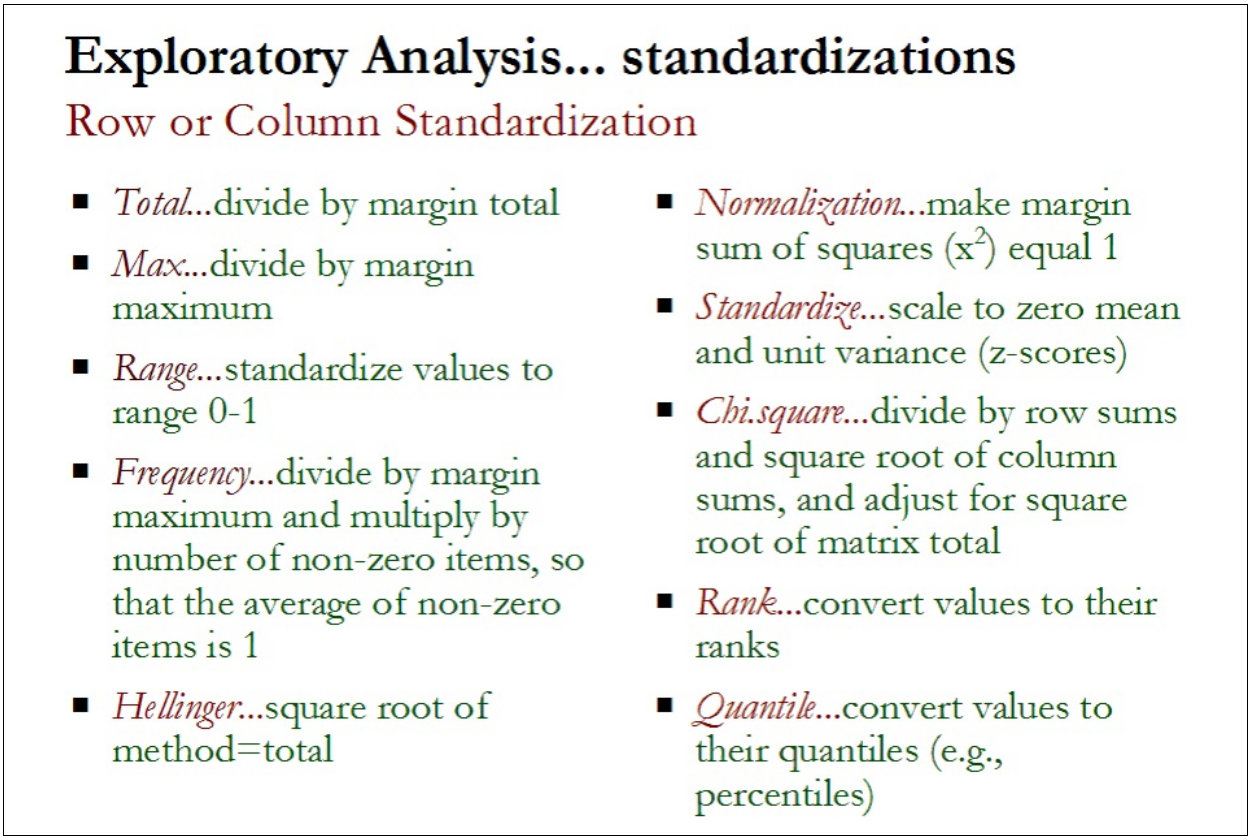



The choice between a column or row standardization has important implications.

A column standardization is appropriate when the principal concern is to adjust for differences (e.g., variances, total abundance) among variables (e.g., species) in order to place them on equal footing, for example when the focus is on the profile *across* sample units.

A row standardization is appropriate when the principal concern is to adjust for differences (e.g., total abundance, diversity) among sample units in order to place them on equal footing, for example when the focus is on the profile *within* a sample unit.

The choice between these two is often confusing and takes careful thought, lots of practice, and lots of trial and error.




There are lots of different common standardizations and we will make no attempt to describe them in detail or illustrate by example their differences, else we would need an entire lecture devoted to this topic. The following list provides a glimpse into the myriad standardizations available, where margin equals either column or row:

- *Total*...divide by margin total
- *Max*...divide by margin maximum
- *Range*...standardize values to range 0-1
- *Frequency*...divide by margin maximum and multiply by number of non-zero items, so that the average of non-zero items is 1
- *Hellinger*...square root of method=total
- *Normalization*...make margin sum of squares (i.e., x^2) equal 1
- *Standardize*...scale to zero mean and unit variance (*z*-scores)
- *Chi.square*...divide by row sums and square root of column sums, and adjust for square root of matrix total
- *Rank*...convert values to their ranks
- *Quantile*...convert values to their quantiles (e.g., percentiles)

*Some rules of thumb*

Here are some general rules of thumb for using standardizations:

- The effect of standardization on the analysis depends on the variability among rows and/or columns. If these values are small, say <50, it is unlikely that standardization will accomplish much. However, if these values are large, say >100, then it is likely that standardization will have a large effect on the results.
- Consider using row standardizations for species data sets, commonly row normalization, chi.square, total and hellinger standardizations based on recommendations in Legendre and Gallagher 2001.
- Consider column standardizations to "equalize" variables measured in different units and scales, commonly column z-scores, normalization, total, and range standardizations.

## Exploratory Analysis... standardizations

Some Rules of Thumb

- Effect of standardization on analysis depends on variability among rows and/or columns

Table 9.2 (McCune and Grace 2002). Evaluation of degree of variability in row or column totals as measured with the coefficient of variation of row or column totals.

| CV (%) | Variability among rows or column |
|---|---|
| <50 | Small. Relativization usually has small effect on qualitative outcome of the analysis |
| 50-100 | Moderate (with a corresponding moderate effect on the outcome of further analysis) |
| 100-300 | Large. Large effect on results |
| >300 | Very large |

## Exploratory Analysis... standardizations

Some Rules of Thumb

*Raw Data Matrix*

| Site | A | B | C | D | E | F | Total |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 3 | 3 | 1 | 10 |
| 2 | 2 | 2 | 4 | 6 | 6 | 0 | 20 |
| 3 | 10 | 10 | 20 | 30 | 30 | 0 | 100 |
| 4 | 3 | 3 | 2 | 1 | 1 | 0 | 10 |
| 5 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| 6 | 0 | 0 | 0 | 0 | 20 | 0 | 20 |
| Total | 16 | 16 | 27 | 40 | 61 | 1 | 161 |

*Row Total Standardization*

| Site | A | B | C | D | E | F | Total |
|---|---|---|---|---|---|---|---|
| 1 | 0.10 | 0.10 | 0.10 | 0.30 | 0.30 | 0.10 | 1 |
| 2 | 0.10 | 0.10 | 0.20 | 0.30 | 0.30 | 0.00 | 1 |
| 3 | 0.10 | 0.10 | 0.20 | 0.30 | 0.30 | 0.00 | 1 |
| 4 | 0.30 | 0.30 | 0.20 | 0.10 | 0.10 | 0.00 | 1 |
| 5 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 | 0.00 | 1 |
| 6 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 | 0.00 | 1 |
| Total | 0.60 | 0.60 | 0.70 | 1.00 | 3.00 | 0.10 | 6 |

- Consider row standardizations for species data sets, commonly:
  - Row *normalize* (Euclidean distance (ED) = chord distance)

(From Legendre and Gallagher 2001)

  - Row *chi.square* (ED = chi.square distance of CA/CCA)
  - Row *total* (ED = species profile distance)
  - Row *hellinger* (ED = Hellinger distance)

## Exploratory Analysis... standardizations

Some Rules of Thumb

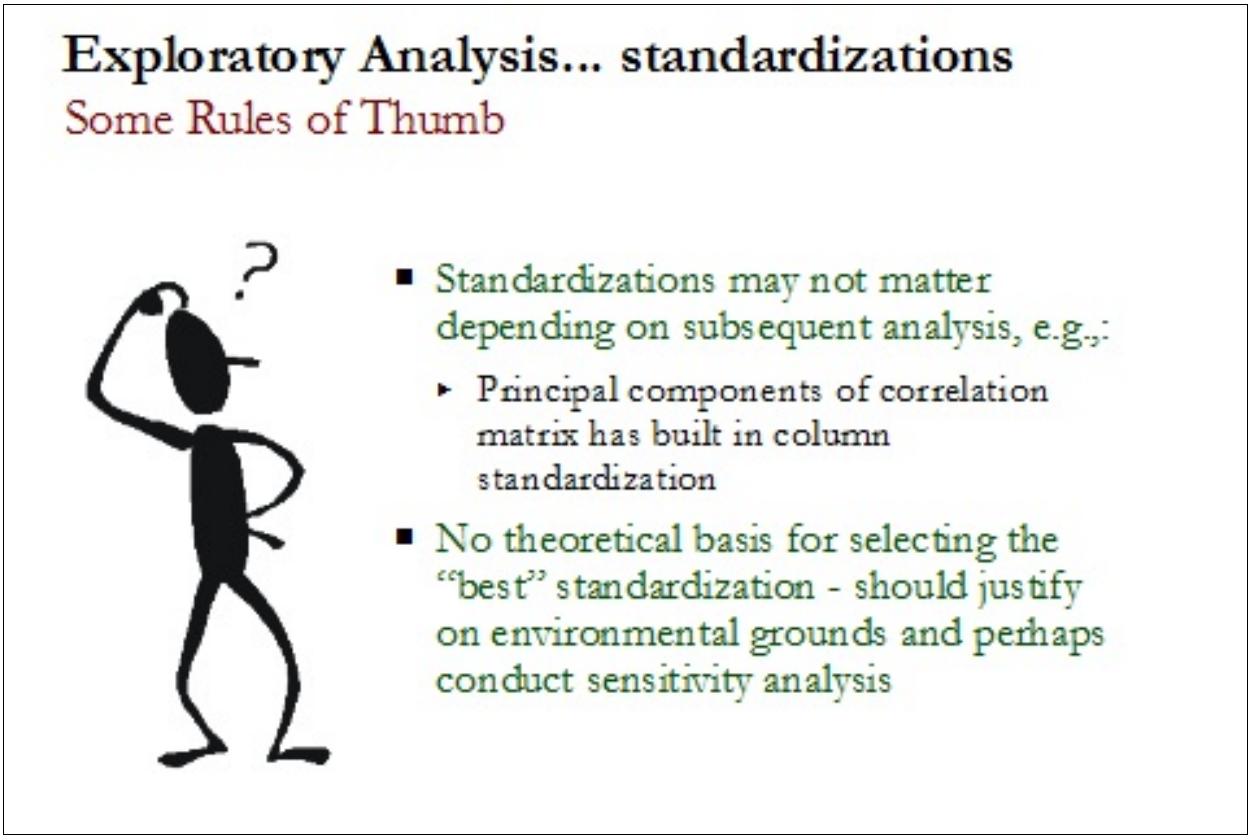
*Raw Data Matrix*

| Site | A | B | C | D | E | F | Total |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 3 | 3 | 1 | 10 |
| 2 | 2 | 2 | 4 | 6 | 6 | 0 | 20 |
| 3 | 10 | 10 | 20 | 30 | 30 | 0 | 100 |
| 4 | 3 | 3 | 2 | 1 | 1 | 0 | 10 |
| 5 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| 6 | 0 | 0 | 0 | 0 | 20 | 0 | 20 |
| Total | 16 | 16 | 27 | 40 | 61 | 1 | 161 |

$b_{ij}=(x_{ij}-\bar{x}_j)/s_j$

| Site | A | B | C | D | E | F | Total |
|---|---|---|---|---|---|---|---|
| 1 | -0.44 | -0.44 | -0.45 | -0.31 | -0.59 | 2.04 | -0.20 |
| 2 | -0.18 | -0.18 | -0.06 | -0.06 | -0.35 | -0.41 | -1.23 |
| 3 | 1.94 | 1.94 | 2.00 | 2.00 | 1.64 | -0.41 | 9.12 |
| 4 | 0.09 | 0.09 | -0.32 | -0.49 | -0.76 | -0.41 | -1.80 |
| 5 | -0.71 | -0.71 | -0.58 | -0.57 | -0.76 | -0.41 | -3.73 |
| 6 | -0.71 | -0.71 | -0.58 | -0.57 | 0.82 | -0.41 | -2.16 |
| Total | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

- Consider column standardizations to "equalize" variables measured in different units and scales, commonly:
  - Column *standardize* (z-scores = zero mean and unit variance)
  - Column *normalize* (uncentered with unit variance)
  - Column *total* (col sums = 1)
  - Column *range* (col range 0-1)

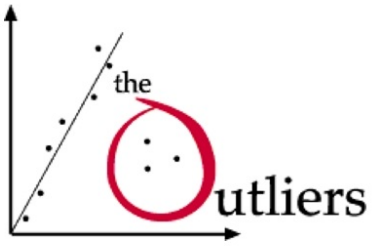


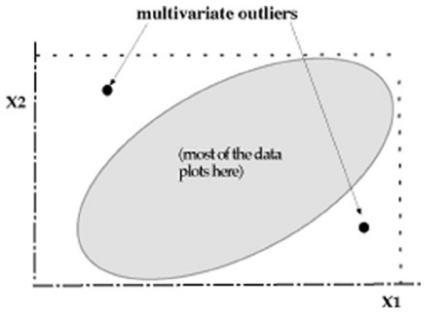

- Standardizations may not matter depending on the subsequent statistical analysis employed. For example, column standardization is not necessary for analyses that use the variables one at a time (e.g., ordination overlays) or for analyses with built-in standardization (e.g., principal components analysis of a correlation matrix).
- Before applying any standardization, be sure to understand what the standardization does. In same cases, standardization is built into the subsequent analyses and therefore unnecessary – but to know this requires that we already understand the mechanics of the methods we intend to use (which we haven't gotten to yet). At this point, we might simply explore what various standardizations do to data so that we are ready and able to standardize the data as needed when we decide on a particular statistical procedure.
- Ultimately, I'm not sure that there is any theoretical basis for selecting the "best" standardization - we should justify our choice on environmental grounds and perhaps conduct sensitivity analysis (i.e., examine how changing the standardization method or whether to standardize or not effects the results).

## Exploratory Analysis... extreme values

"Outliers"

- What are outliers?
  - Sample units with extreme values for individual variables (univariate outliers) or sample units with unusual combination of values for more than one variable (mulitvariate outliers)
- Why worry about outliers?
  - Outliers can have a large effect on the outcome of an analysis and therefore can lead to erroneous conclusions




### 9. Extreme values ("outliers")

Environmental data commonly contain values that are "extreme", or considered to be "outliers" in the sense that they are much larger or smaller than the rest of the data and thus fall "outside" the bulk of the data. These so-called outliers can have a large effect on the outcome of an analysis and therefore can lead to erroneous conclusions if not dealt with properly. What constitutes a true "outlier" depends on the question being asked and the analysis being conducted. There is no general rule for deciding whether extreme observations should be considered "outliers" and deleted from the data set before proceeding with the analysis. Nevertheless, it is important to have an understanding of the number and pattern of extreme observations in order to gauge the robustness of the results. A good practice is to repeat the analyses with and without the suspect points and determine if the results are sensitive or robust to their inclusion. If the results are sensitive to the inclusion of these high-leverage points, you should probably carefully consider whether those points represent a meaningful environmental condition, and act on them accordingly.

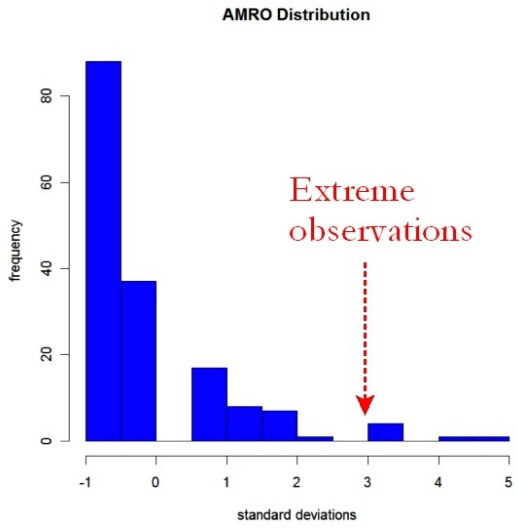
# Exploratory Analysis... extreme values

## "Outliers"

- Univariate outliers:
  - Values say >3 standard deviations from the mean of the variable

Standard deviation scores >3

| | AMGO | AMRO | BAEA | BCCH |
|---|---|---|---|---|
| 81 | 6.50 | 4.91 | NA | NA |
| 82 | 6.50 | NA | NA | NA |
| 83 | 4.27 | 4.30 | NA | 4.29 |
| 84 | 6.50 | NA | NA | NA |
| 85 | NA | NA | NA | 5.44 |
| 87 | NA | NA | NA | NA |
| 89 | NA | NA | NA | 5.44 |
| 90 | NA | NA | NA | NA |
| 91 | NA | NA | 12.73 | NA |




*Univariate outliers*

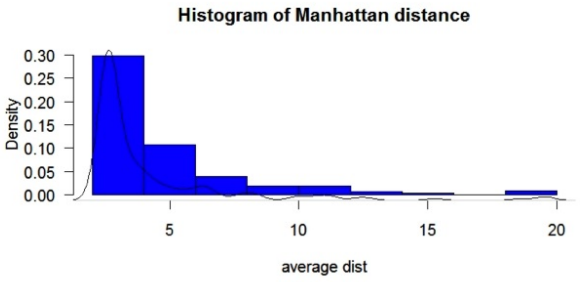
There are several ways to identify extreme values, including both univariate and multivariate methods. It is always a good idea to begin with a univariate inspection. The most common univariate method involves computing the $z$-score standardized values for each variable and looking for values that are greater than say 3 standard deviations from the mean. These are observations that fall outside 99.7% of the data under the assumption of a normal distribution and are regardless of distribution likely to be extreme in the sense of falling outside the bulk of the data.

In the example shown here, there are several species with relative abundance values that are greater than 3 standard deviations from the mean relative abundance of the species, as depicted in the table. Note, these extreme observations show up the histogram of the $z$-scores for the American Robin (AMRO).

# Exploratory Analysis... extreme values

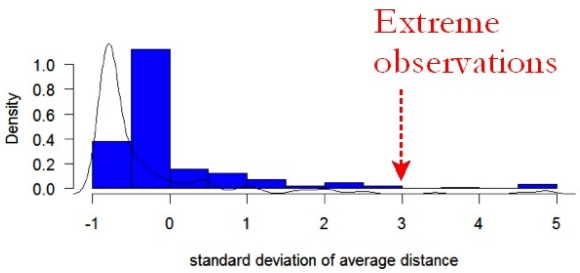
## "Outliers"

- Multivariate outliers:
  - Values say >3 standard deviations in average distance to other samples from the mean average distance among samples




Standard deviation scores >3

| | avedist | sd |
|---|---|---|
| 81 | 18.399 | 4.624 |
| 82 | 19.294 | 4.931 |
| 83 | 19.331 | 4.944 |
| 85 | 15.294 | 3.558 |




*Multivariate outliers*

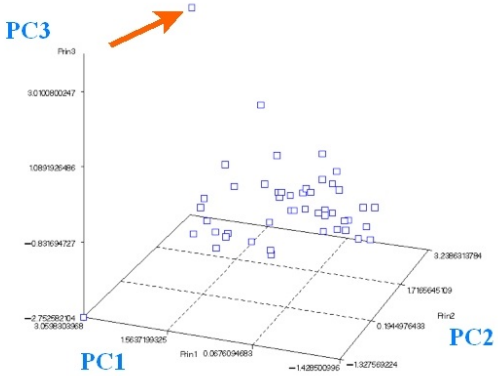
In the context of a multivariate data set, just because an observation is extreme on a single variable, doesn't mean it is going to be a multivariate outlier. More importantly, an observation may not be a univariate outlier and yet still be an outlier when two or more variables are considered jointly. Thus, with multivariate data it is instructive to see if each observation is extreme in multivariate space.

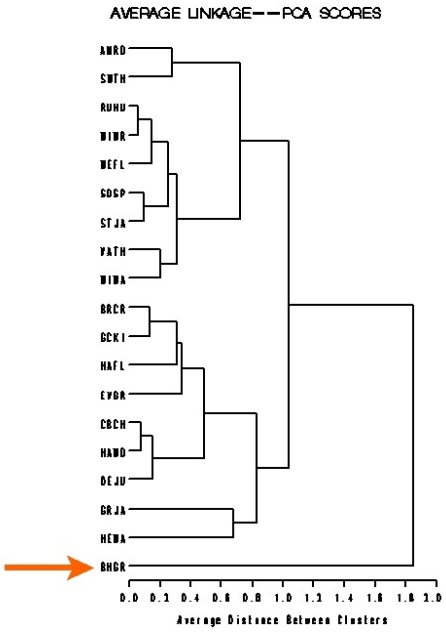
One method for evaluating multivariate outliers is to measure the distance from each sample point to every other sample point based on some measure of environmental distance – unfortunately this is a topic that we do not have time to cover so you will have to take this one on faith, but briefly, if you recall Euclidean distance from basic match (remember the Pythagorean theorem for measuring the distance between points in 2-dimensional space), a simple measure of distance is the Euclidean (or straight line) distance between points in multidimensional space. Then, we compute the average distance from each point to every other point. Points that are extreme in a multivariate sense should have a large average distance to all other points. Finally, we compute the z-scores for the average distances, which simply puts the average distance information on a scale that we can all understand and interpret; i.e., standard deviation units. So, a point that is greater than say 3 standard deviation units from the average in its average multivariate distance is an extreme point and warrants attention. In the example shown here, 4 sample points where identified as being extreme based on a particular distance metric known as Manhattan distance. The histogram merely shows the same thing graphically.

# Exploratory Analysis... extreme values

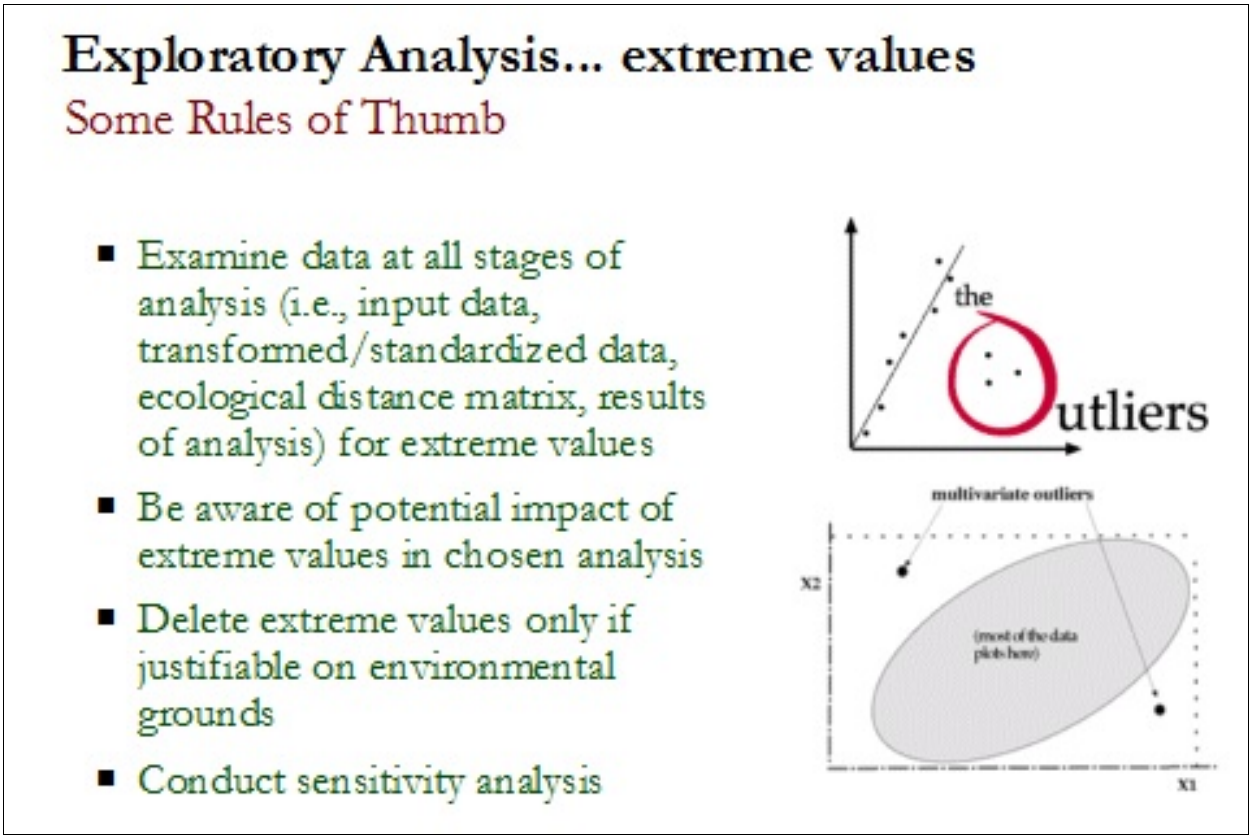
## "Outliers"

- Multivariate outliers:
  - Extreme values usually show up in multivariate plots; e.g., isolated points in ordination plots, single-member clusters in cluster analysis, etc.




Not surprisingly, there are lots of other multivariate techniques that can aid in identifying potential outliers. Shown here are two methods, known as unconstrained ordination and hierarchical cluster analysis. The details of the methods are beyond the scope of this lecture and not important to understanding the concept; they simply reveal the existence of extreme points or outliers in different ways. In the ordination plot on the left, the extreme point shows up as being isolation from all other points in the 3-dimensional scatterplot. In the cluster analysis plot on the right (known as a dendrogram), the extreme point shows up as not connecting (or clustering) to the other points until a much greater environmental distance.




*Some rules of thumb*

Here are some general rules of thumb for dealing with extreme values or outliers:

- Examine data at all stages of analysis (i.e., input data, transformed/standardized data, environmental distance matrix, results of analysis) for extreme values.
- Be aware of the potential impact of extreme values in the chosen statistical analysis.
- Delete extreme values only if justifiable on environmental grounds. As a general rule, observations should not be dropped automatically just because they have extreme values. It is one thing to identify extreme observations that have high leverage on the results, but it is another thing altogether to delete these observations from the data set just because they have high leverage.
- Conduct sensitivity analysis to determine the realize impact of extreme values. Quite simply, this involves conducting the analysis with and with potential outliers and seeing how much the results vary.