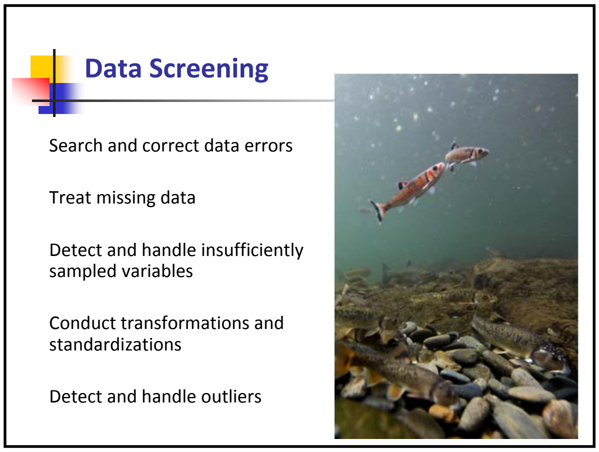
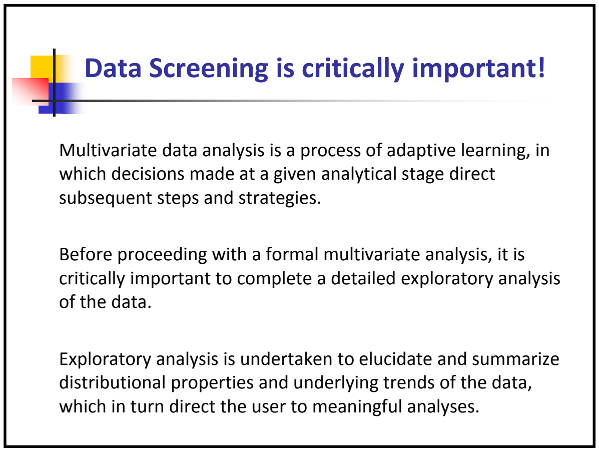


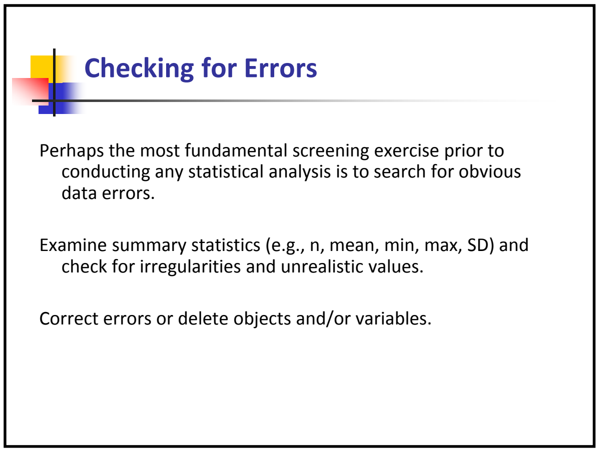
Data Screening

- Search and correct data errors
- Treat missing data
- Detect and handle insufficiently sampled variables
- Conduct transformations and standardizations
- Detect and handle outliers



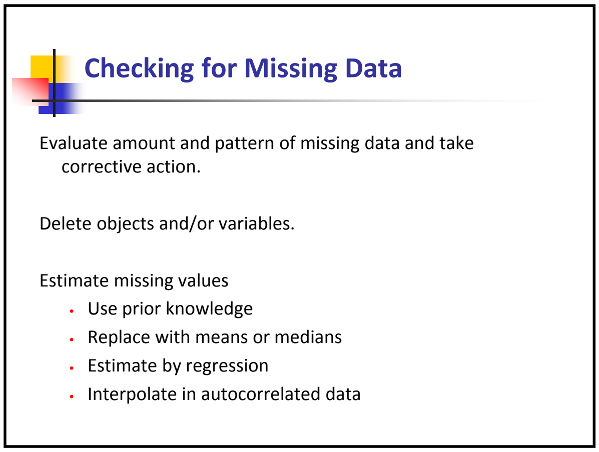
Data Screening is critically important!

- Multivariate data analysis is a process of adaptive learning, in which decisions made at a given analytical stage direct subsequent steps and strategies.
- Before proceeding with a formal multivariate analysis, it is critically important to complete a detailed exploratory analysis of the data.
- Exploratory analysis is undertaken to elucidate and summarize distributional properties and underlying trends of the data, which in turn direct the user to meaningful analyses.



Checking for Errors

- Perhaps the most fundamental screening exercise prior to conducting any statistical analysis is to search for obvious data errors.
- Examine summary statistics (e.g., n, mean, min, max, SD) and check for irregularities and unrealistic values.
- Correct errors or delete objects and/or variables.



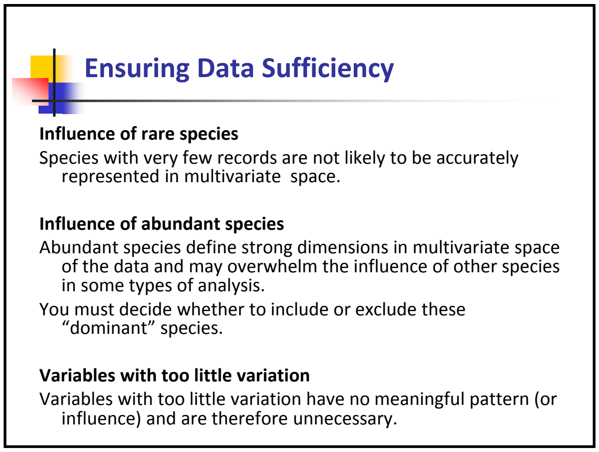
Checking for Missing Data

Evaluate amount and pattern of missing data and take corrective action.

Delete objects and/or variables.

Estimate missing values

- Use prior knowledge
- Replace with means or medians
- Estimate by regression
- Interpolate in autocorrelated data



Ensuring Data Sufficiency

Influence of rare species

Species with very few records are not likely to be accurately represented in multivariate space.

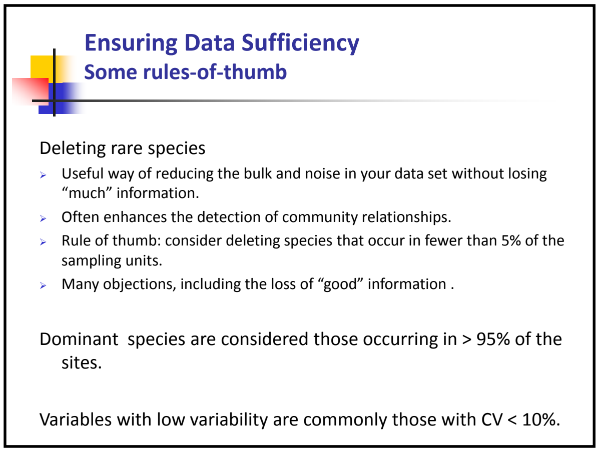
Influence of abundant species

Abundant species define strong dimensions in multivariate space of the data and may overwhelm the influence of other species in some types of analysis.

You must decide whether to include or exclude these "dominant" species.

Variables with too little variation

Variables with too little variation have no meaningful pattern (or influence) and are therefore unnecessary.



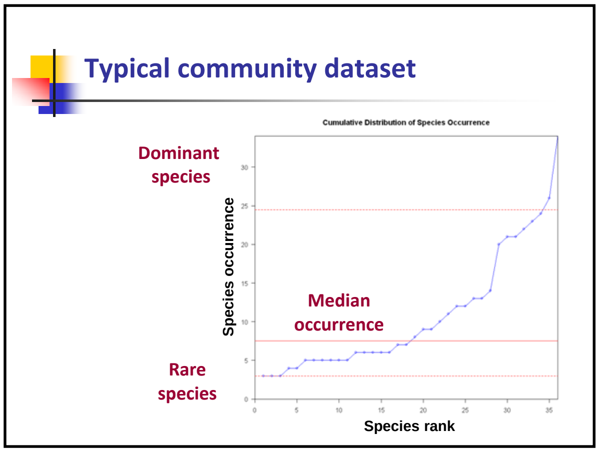
Ensuring Data Sufficiency
Some rules-of-thumb

Deleting rare species

- Useful way of reducing the bulk and noise in your data set without losing "much" information.
- Often enhances the detection of community relationships.
- Rule of thumb: consider deleting species that occur in fewer than 5% of the sampling units.
- Many objections, including the loss of "good" information .

Dominant species are considered those occurring in > 95% of the sites.

Variables with low variability are commonly those with CV < 10%.

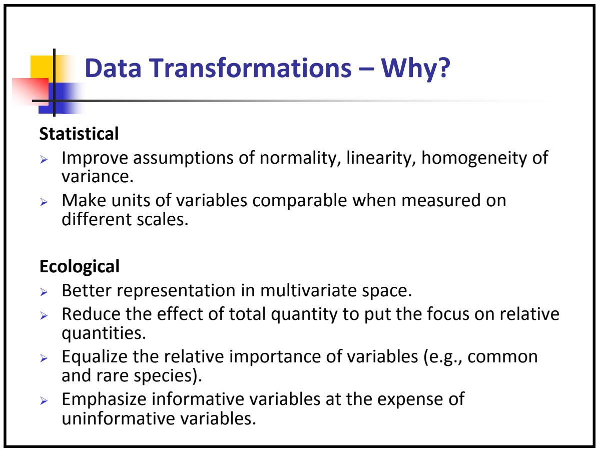


Typical community dataset

Dominant species

Median occurrence

Rare species



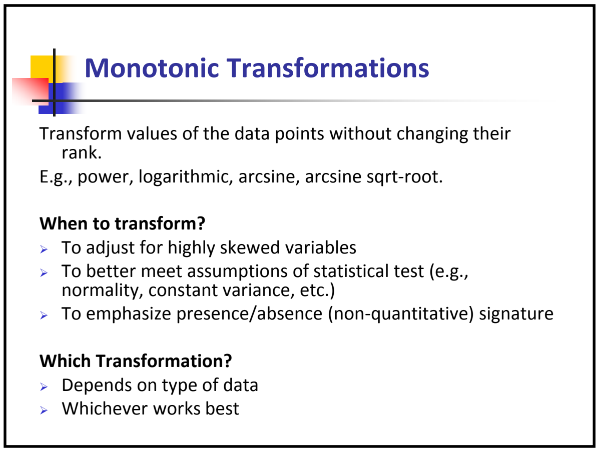
Data Transformations – Why?

Statistical

- Improve assumptions of normality, linearity, homogeneity of variance.
- Make units of variables comparable when measured on different scales.

Ecological

- Better representation in multivariate space.
- Reduce the effect of total quantity to put the focus on relative quantities.
- Equalize the relative importance of variables (e.g., common and rare species).
- Emphasize informative variables at the expense of uninformative variables.



Monotonic Transformations

Transform values of the data points without changing their rank.

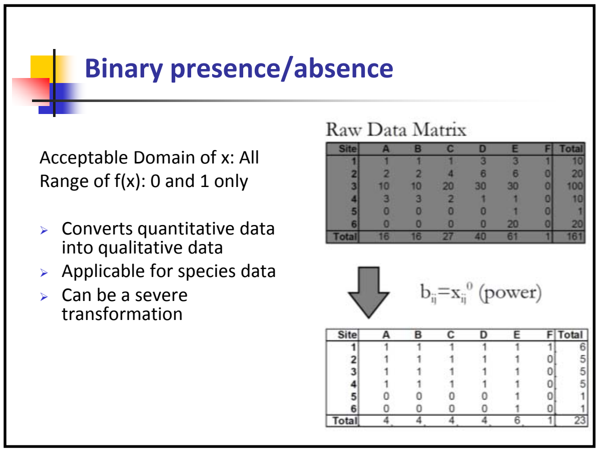
E.g., power, logarithmic, arcsine, arcsine sqrt-root.

When to transform?

- To adjust for highly skewed variables
- To better meet assumptions of statistical test (e.g., normality, constant variance, etc.)
- To emphasize presence/absence (non-quantitative) signature

Which Transformation?

- Depends on type of data
- Whichever works best



Binary presence/absence

Acceptable Domain of x: All
Range of f(x): 0 and 1 only

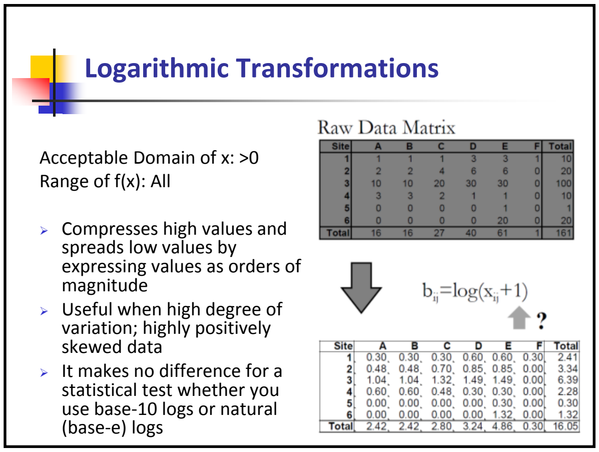
- Converts quantitative data into qualitative data
- Applicable for species data
- Can be a severe transformation

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

$b_{ij}=x_{ij}^0$ (power)

Site	A	B	C	D	E	F	Total
1	1	1	1	1	1	1	6
2	1	1	1	1	1	0	5
3	1	1	1	1	1	0	5
4	1	1	1	1	1	0	5
5	0	0	0	0	1	0	1
6	0	0	0	0	1	0	1
Total	4	4	4	4	6	1	23



Logarithmic Transformations

Acceptable 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; highly positively skewed data
- It makes no difference for a statistical test whether you use base-10 logs or natural (base-e) logs

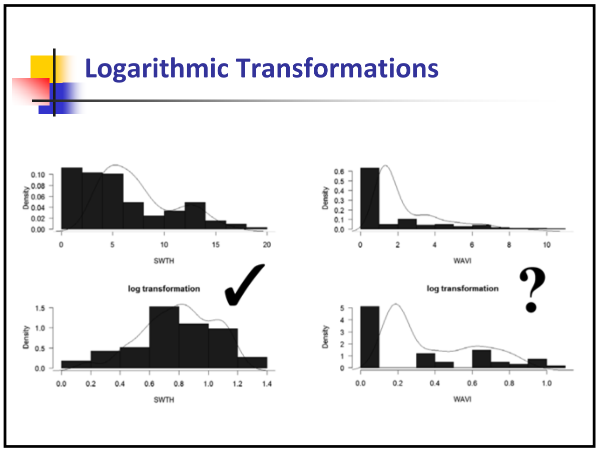
Raw Data Matrix

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

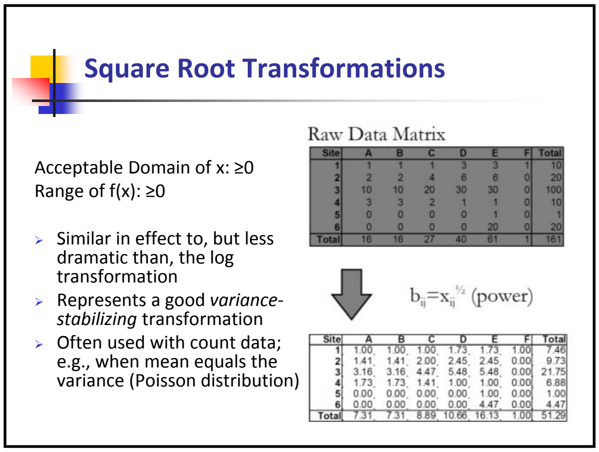
$b_{ij}=\log(x_{ij}+1)$

?

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



Logarithmic Transformations



Square Root Transformations

Acceptable Domain of x: ≥0
Range of f(x): ≥0

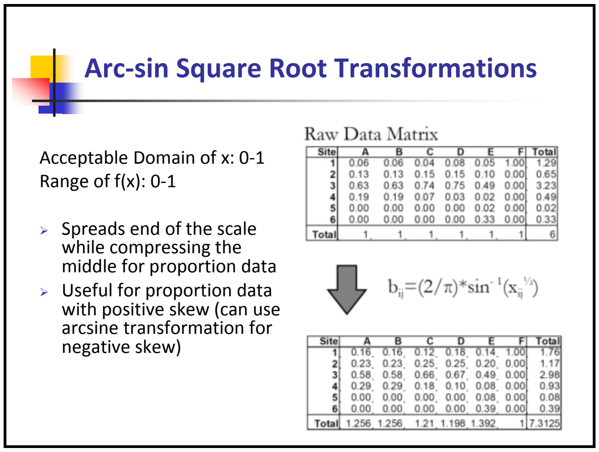
- Similar in effect to, but less dramatic than, the log transformation
- Represents a good *variance-stabilizing* transformation
- Often used with count data; e.g., when mean equals the variance (Poisson distribution)

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

$b_{ij}=x_{ij}^{1/2}$ (power)

Site	A	B	C	D	E	F	Total
1	1.00	1.00	1.00	1.73	1.73	1.00	7.46
2	1.41	1.41	2.00	2.45	2.45	0.00	9.73
3	3.16	3.16	4.47	5.48	5.48	0.00	21.75
4	1.73	1.73	1.41	1.00	1.00	0.00	6.88
5	0.00	0.00	0.00	0.00	1.00	0.00	1.00
6	0.00	0.00	0.00	0.00	4.47	0.00	4.47
Total	7.31	7.31	8.89	10.66	16.13	1.00	51.29



Arc-sin Square Root Transformations

Acceptable 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)

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

Monotonic Transformations
Some Rules-of-thumb

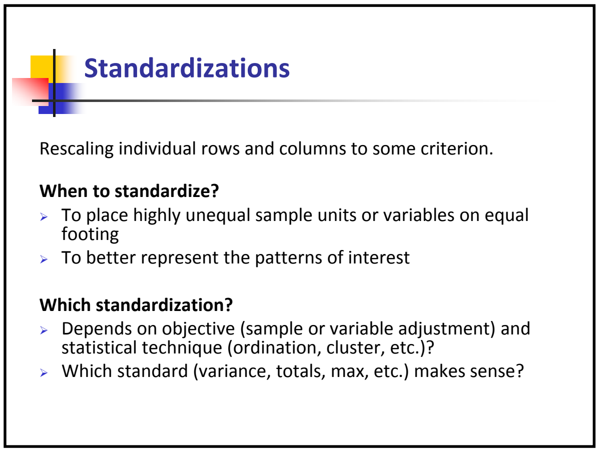
Use a *log* or *square root* transformation for "highly" skewed data or ranging over several (>2) orders of magnitude

Use *arcsine square root* transformation for proportion data

If applied to related variable set (e.g., species), then use same transformation (e.g., log) so that all are scaled the same

Consider binary (presence/absence) transformation when:

- percent of zero values is high (say >50%)
- number of distinct values is low (say < 10)



Standardizations

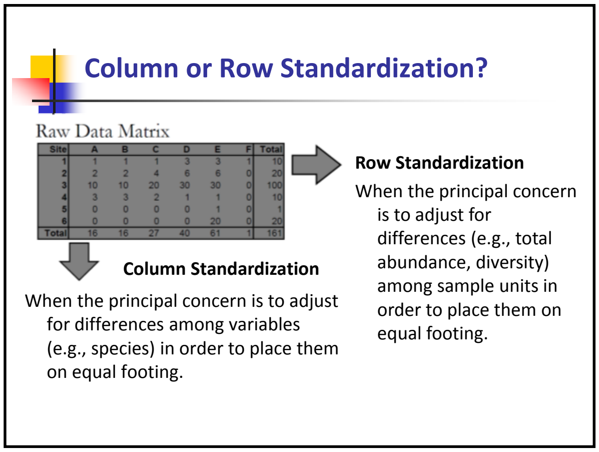
Rescaling individual rows and columns to some criterion.

When to standardize?

- To place highly unequal sample units or variables on equal footing
- To better represent the patterns of interest

Which standardization?

- Depends on objective (sample or variable adjustment) and statistical technique (ordination, cluster, etc.)?
- Which standard (variance, totals, max, etc.) makes sense?



Column or Row Standardization?

Raw Data Matrix

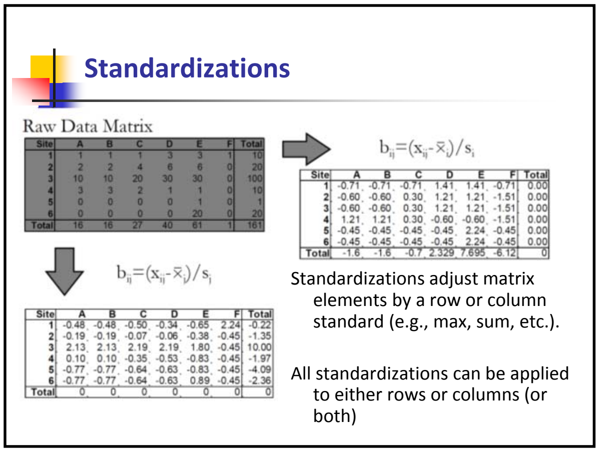
Site	A	B	C	D	E	F	Total
1	1	1	1	3	3	1	10
2	2	2	4	6	6	0	20
3	10	10	20	30	30	0	100
4	3	3	2	1	1	0	10
5	0	0	0	0	1	0	1
6	0	0	0	0	20	0	20
Total	16	16	27	40	61	1	161

Row Standardization

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.

Column Standardization

When the principal concern is to adjust for differences among variables (e.g., species) in order to place them on equal footing.



Standardizations

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

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

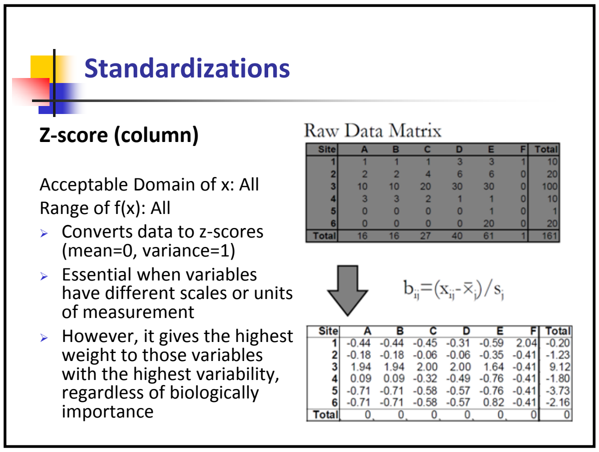
Site	A	B	C	D	E	F	Total
1	-0.71	-0.71	-0.71	1.41	1.41	-0.71	0.00
2	-0.60	-0.60	0.30	1.21	1.21	-1.51	0.00
3	-0.60	-0.60	0.30	1.21	1.21	-1.51	0.00
4	1.21	1.21	0.30	-0.60	-0.60	-1.51	0.00
5	-0.45	-0.45	-0.45	-0.45	2.24	-0.45	0.00
6	-0.45	-0.45	-0.45	-0.45	2.24	-0.45	0.00
Total	-1.6	-1.6	-0.7	2.329	7.695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



Standardizations

Z-score (column)

Acceptable Domain of x: All
Range of f(x): All

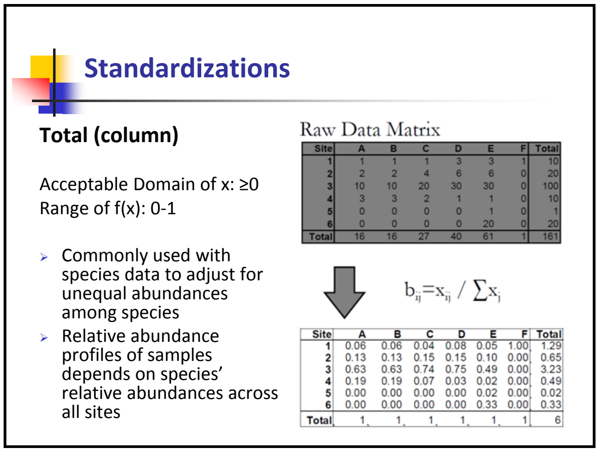
- Converts data to z-scores (mean=0, variance=1)
- Essential when variables have different scales or units of measurement
- However, it gives the highest weight to those variables with the highest variability, regardless of biologically importance

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



Standardizations

Total (column)

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

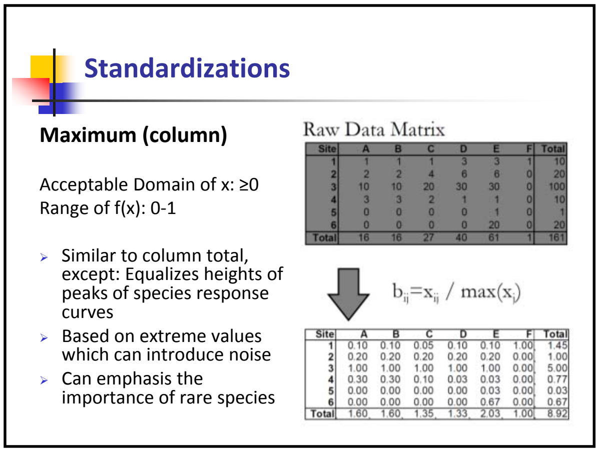
- Commonly used with species data to adjust for unequal abundances among species
- Relative abundance profiles of samples depends on species' relative abundances across all sites

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

$b_{ij}=x_{ij} / \sum x_j$

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



Standardizations

Maximum (column)

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

- Similar to column total, except: Equalizes heights of peaks of species response curves
- Based on extreme values which can introduce noise
- Can emphasis the importance of rare species

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

$b_{ij}=x_{ij} / \max(x_j)$

Site	A	B	C	D	E	F	Total
1	0.10	0.10	0.05	0.10	0.10	1.00	1.45
2	0.20	0.20	0.20	0.20	0.20	0.00	1.00
3	1.00	1.00	1.00	1.00	1.00	0.00	5.00
4	0.30	0.30	0.10	0.03	0.03	0.00	0.77
5	0.00	0.00	0.00	0.00	0.03	0.00	0.03
6	0.00	0.00	0.00	0.00	0.67	0.00	0.67
Total	1.60	1.60	1.35	1.33	2.03	1.00	8.92

Standardizations
Some Rules-of-thumb

The effect of data standardization on your results will depend on the amount of variability among rows and/or columns.

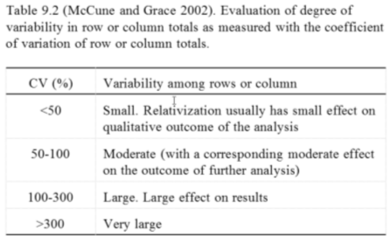


Table 9.2 (McCune and Grace 2002). Evaluation of degree of variability in row or column totals as measured with the coefficient of variation of row or column totals.

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

Standardizations
Some Rules-of-thumb

- Consider row standardizations for species datasets , e.g., relative abundance within sites
- Consider column standardizations to “equalize” variables measured in different units and scales. Commonly used standardizations include:
  - Standardize using z-scores
  - Normalize using column totals

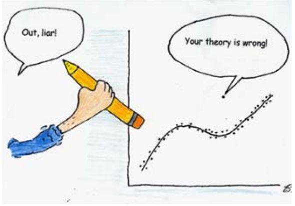
Standardizations
Some Rules-of-thumb

- Standardizations may not matter depending on analysis
  - Principal components of correlation matrix has built in column standardization
  - Correspondence analysis of species data set has essentially a built in chi-square standardization
- There is no theoretical basis for selecting the “best” standardization – you must justify your decision on biological grounds (and perhaps conduct sensitivity analysis)
- Consult: Legendre and Gallagher. 2001. Ecologically meaningful transformations for ordination of species data. Oecologia 129: 271-280.

Data Screening for Outliers

Outliers are sample units with extreme values for individual variables (univariate outliers) or sample units with unusual combination of values for >2 variables (multivariate outliers).

Outliers can have a large effect on the outcome of an analysis and therefore can lead to erroneous conclusions.



Data Screening for Outliers

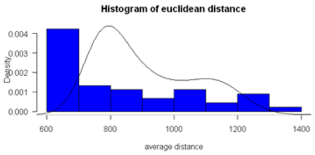
Univariate outliers are "easy" to detect, but multivariate outliers are more tricky.

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.

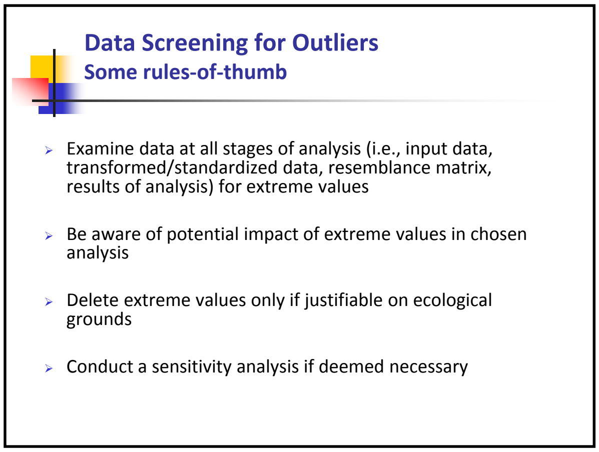
More importantly, an observation may not be a univariate outlier and yet still be an outlier when two or more variables are considered jointly.

Data Screening for Outliers

To do this, it is recommended to compute the average distance of each sample to all other samples using an appropriate distance measure – ideally the distance measure to be used in subsequent analyses.

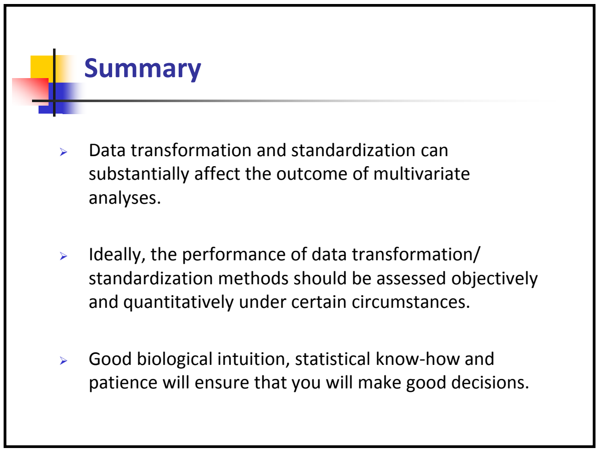


Also, you can examine the results of subsequent analyses for extreme values (e.g., isolated points in ordination plots, single-member clusters in cluster analysis, etc.)



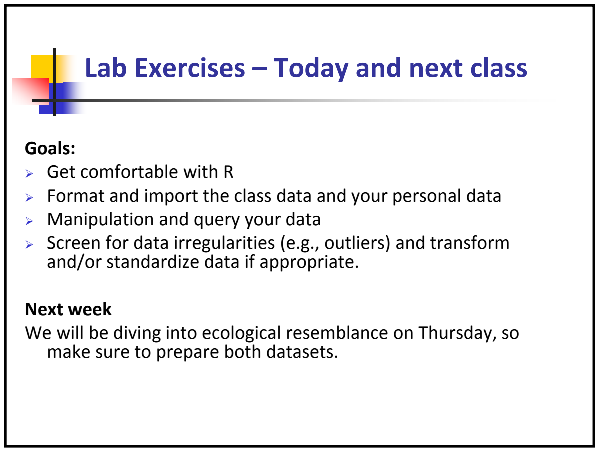
Data Screening for Outliers
Some rules-of-thumb

- Examine data at all stages of analysis (i.e., input data, transformed/standardized data, resemblance matrix, results of analysis) for extreme values
- Be aware of potential impact of extreme values in chosen analysis
- Delete extreme values only if justifiable on ecological grounds
- Conduct a sensitivity analysis if deemed necessary



Summary

- Data transformation and standardization can substantially affect the outcome of multivariate analyses.
- Ideally, the performance of data transformation/ standardization methods should be assessed objectively and quantitatively under certain circumstances.
- Good biological intuition, statistical know-how and patience will ensure that you will make good decisions.



Lab Exercises – Today and next class

Goals:

- Get comfortable with R
- Format and import the class data and your personal data
- Manipulation and query your data
- Screen for data irregularities (e.g., outliers) and transform and/or standardize data if appropriate.

Next week

We will be diving into ecological resemblance on Thursday, so make sure to prepare both datasets.