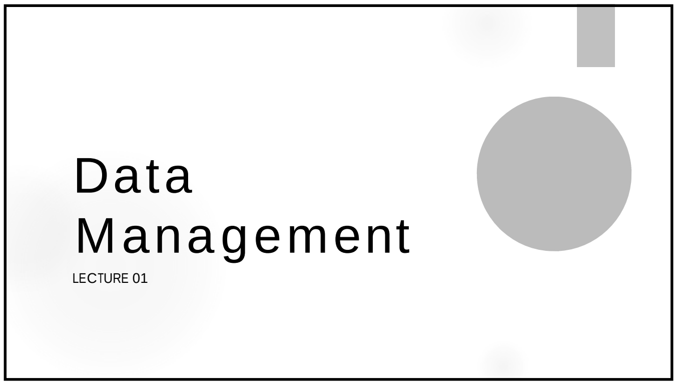
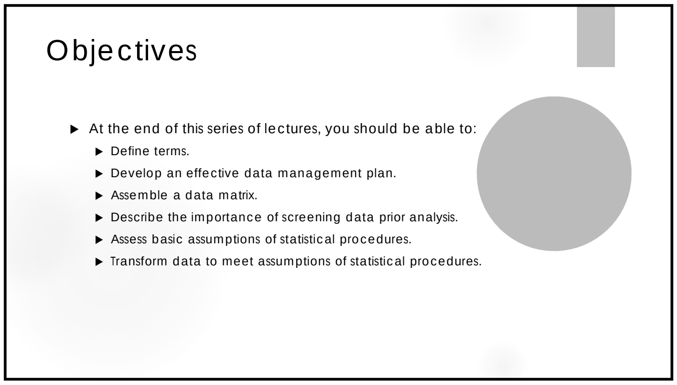


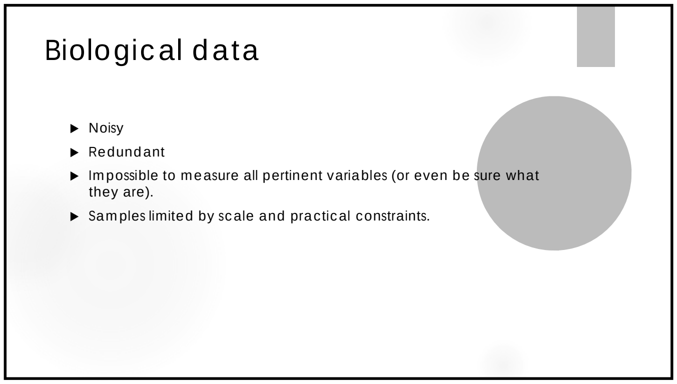
Data Management

LECTURE 01



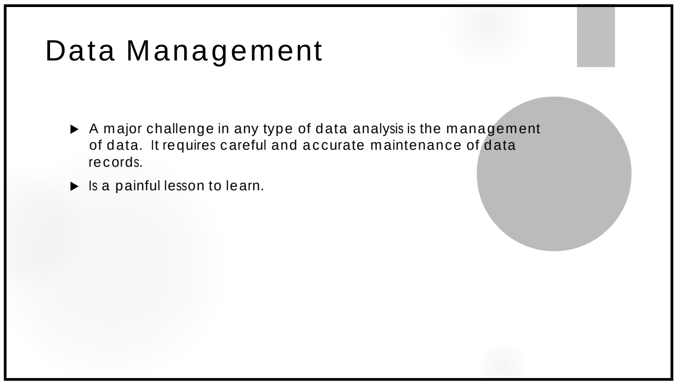
Objectives

- At the end of this series of lectures, you should be able to:
  - Define terms.
  - Develop an effective data management plan.
  - Assemble a data matrix.
  - Describe the importance of screening data prior analysis.
  - Assess basic assumptions of statistical procedures.
  - Transform data to meet assumptions of statistical procedures.



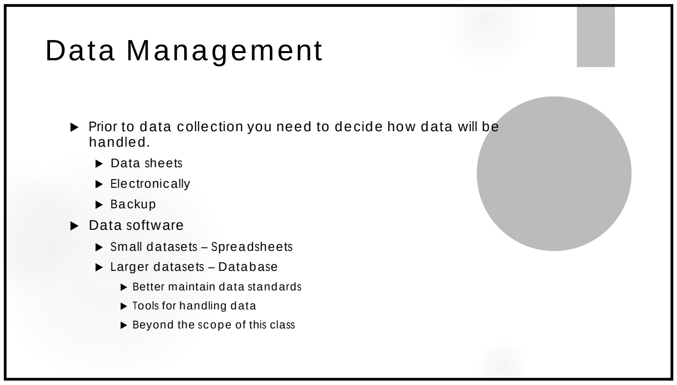
Biological data

- Noisy
- Redundant
- Impossible to measure all pertinent variables (or even be sure what they are).
- Samples limited by scale and practical constraints.



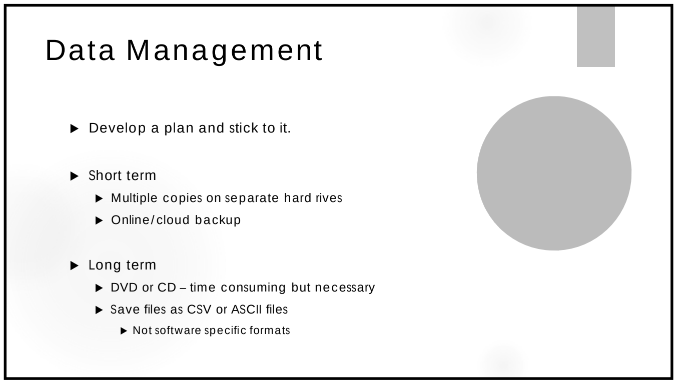
Data Management

- A major challenge in any type of data analysis is the management of data. It requires careful and accurate maintenance of data records.
- Is a painful lesson to learn.



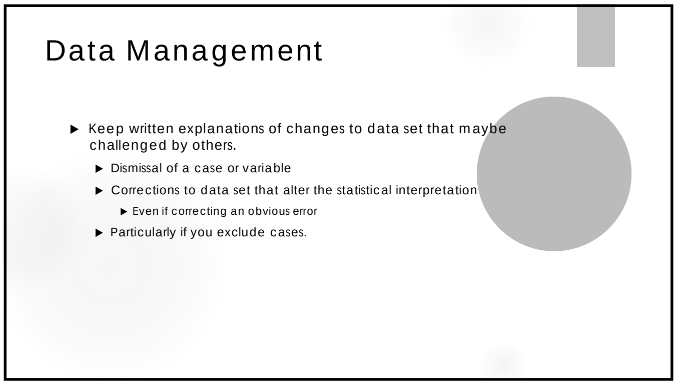
Data Management

- Prior to data collection you need to decide how data will be handled.
  - Data sheets
  - Electronically
  - Backup
- Data software
  - Small datasets – Spreadsheets
  - Larger datasets – Database
    - Better maintain data standards
    - Tools for handling data
    - Beyond the scope of this class



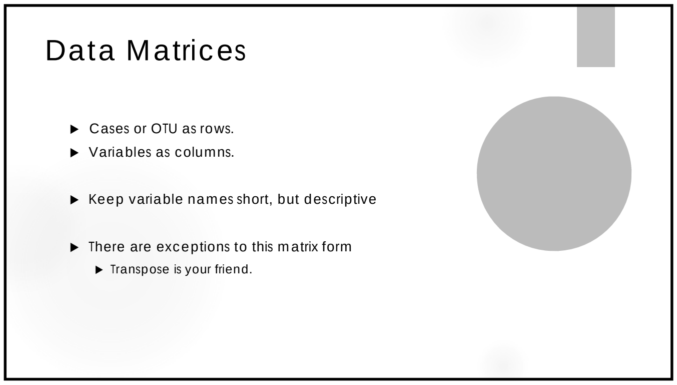
Data Management

- Develop a plan and stick to it.
- Short term
  - Multiple copies on separate hard rives
  - Online/cloud backup
- Long term
  - DVD or CD – time consuming but necessary
  - Save files as CSV or ASCII files
    - Not software specific formats



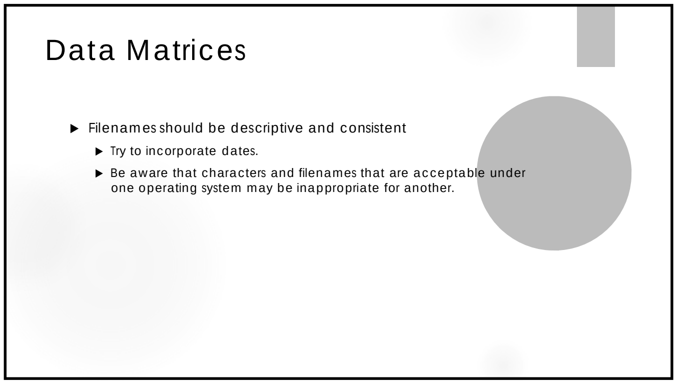
Data Management

- Keep written explanations of changes to data set that maybe challenged by others.
  - Dismissal of a case or variable
  - Corrections to data set that alter the statistical interpretation
    - Even if correcting an obvious error
  - Particularly if you exclude cases.



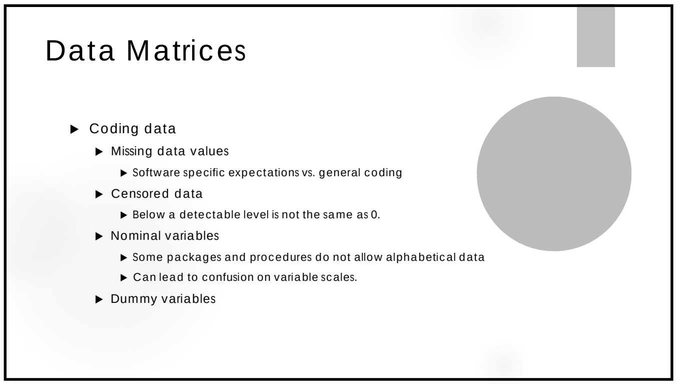
Data Matrices

- Cases or OTU as rows.
- Variables as columns.
- Keep variable names short, but descriptive
- There are exceptions to this matrix form
  - Transpose is your friend.



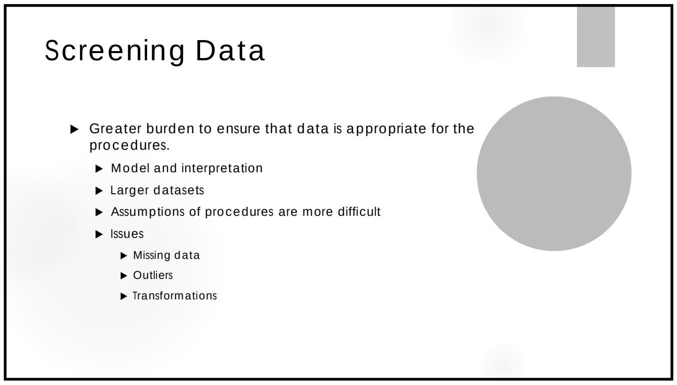
Data Matrices

- Filenames should be descriptive and consistent
  - Try to incorporate dates.
  - Be aware that characters and filenames that are acceptable under one operating system may be inappropriate for another.



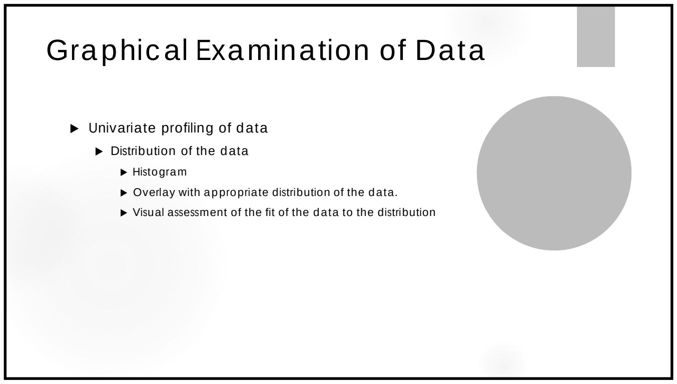
Data Matrices

- Coding data
  - Missing data values
    - Software specific expectations vs. general coding
  - Censored data
    - Below a detectable level is not the same as 0.
  - Nominal variables
    - Some packages and procedures do not allow alphabetical data
    - Can lead to confusion on variable scales.
  - Dummy variables



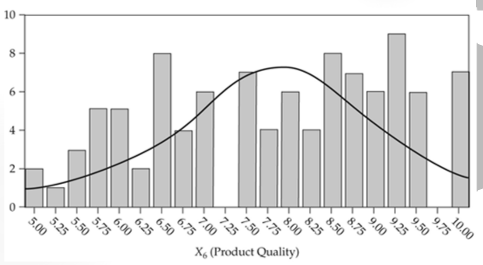
Screening Data

- Greater burden to ensure that data is appropriate for the procedures.
  - Model and interpretation
  - Larger datasets
  - Assumptions of procedures are more difficult
  - Issues
    - Missing data
    - Outliers
    - Transformations



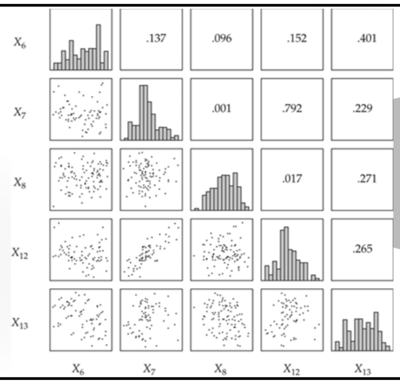
Graphical Examination of Data

- Univariate profiling of data
  - Distribution of the data
    - Histogram
    - Overlay with appropriate distribution of the data.
    - Visual assessment of the fit of the data to the distribution



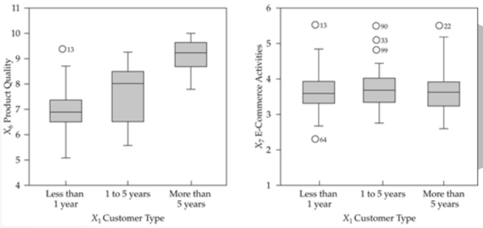
Graphical Examination of Data

- Bivariate profiling of data
  - Scatterplot
    - Useful in assessing correlations and regressions
      - Direction of relationship (positive or negative)
      - Strength of the relationship (r2)
      - Nature of relationship (linear or nonlinear)
      - Lattice for a complete dataset



Graphical Examination of Data

- Bivariate profiling: Group differences
  - Boxplot
    - T-test or ANOVA for groups
    - Identify outliers

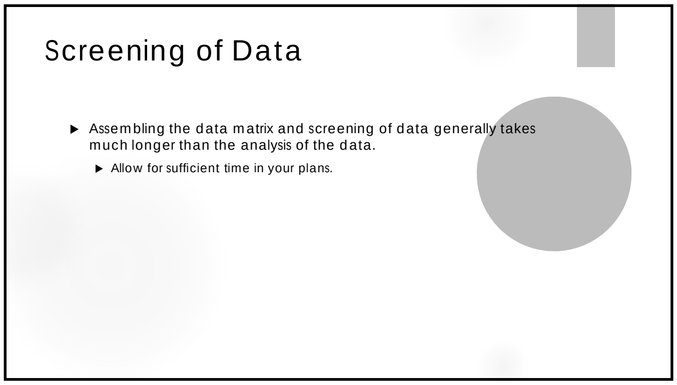


Graphical Examination of Data

- Multivariate Profiling
  - Weird and rarely used

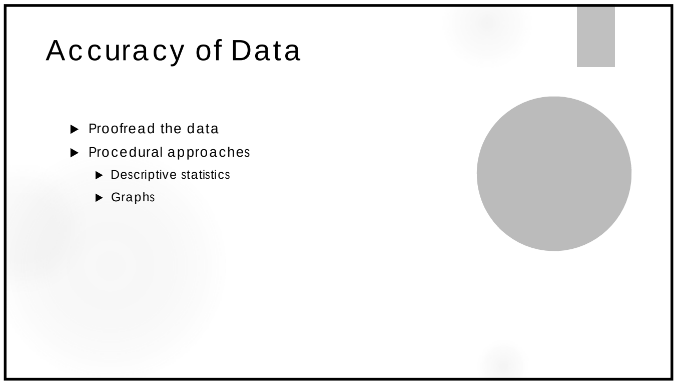
- Profiling techniques do not replace analyses, but increase confidence that the relationships are real.

- IMPORTANT: Use graphs and statistics to develop the strongest case that you can that your interpretation is valid.



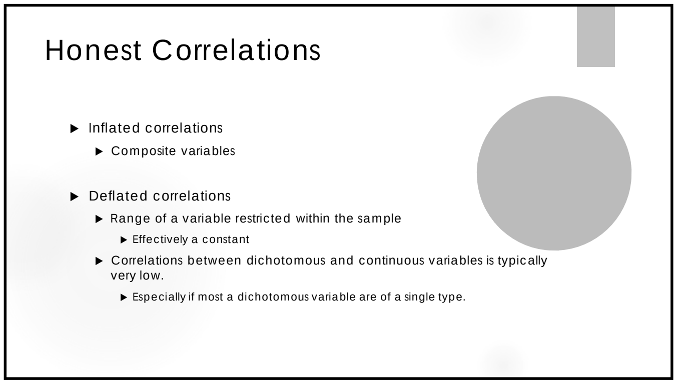
Screening of Data

- Assembling the data matrix and screening of data generally takes much longer than the analysis of the data.
  - Allow for sufficient time in your plans.



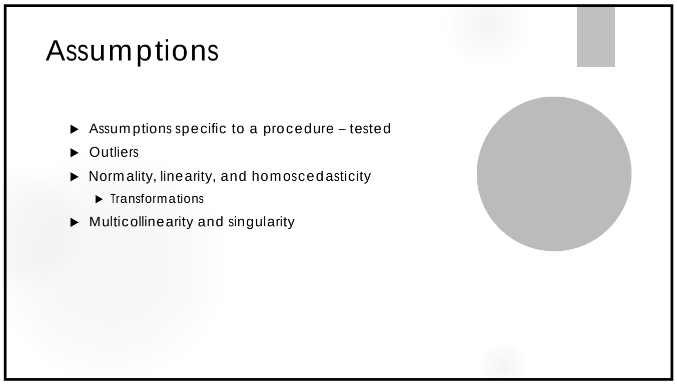
Accuracy of Data

- Proofread the data
- Procedural approaches
  - Descriptive statistics
  - Graphs



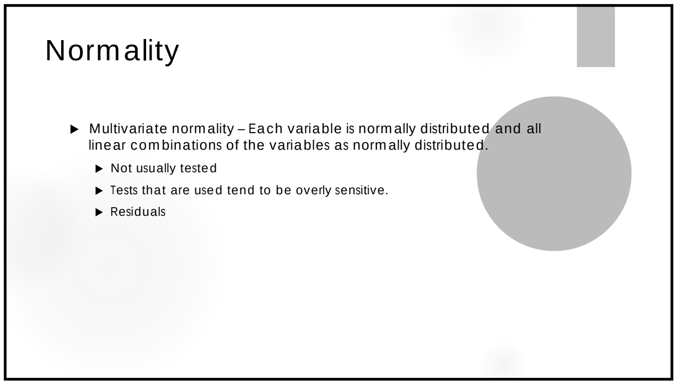
Honest Correlations

- Inflated correlations
  - Composite variables
- Deflated correlations
  - Range of a variable restricted within the sample
    - Effectively a constant
  - Correlations between dichotomous and continuous variables is typically very low.
    - Especially if most a dichotomous variable are of a single type.



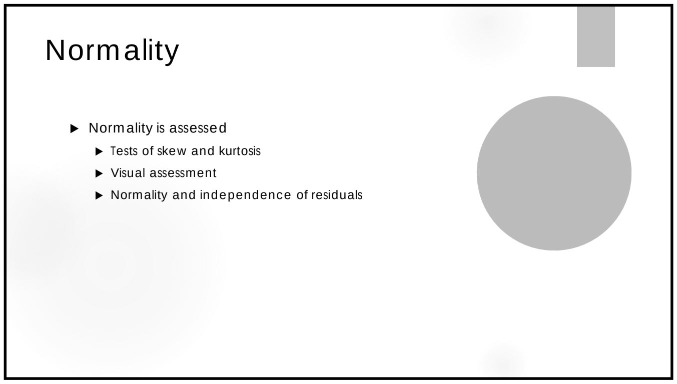
Assumptions

- Assumptions specific to a procedure – tested
- Outliers
- Normality, linearity, and homoscedasticity
  - Transformations
- Multicollinearity and singularity



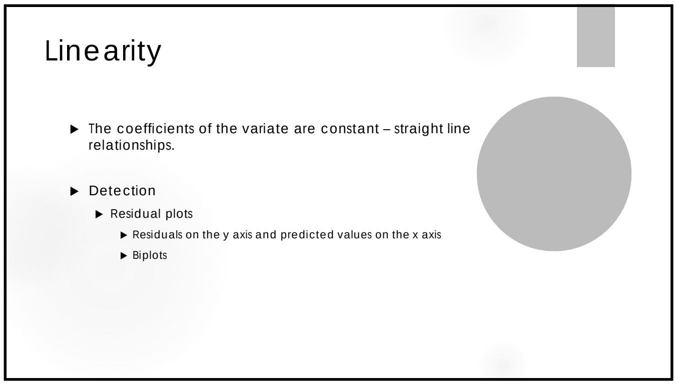
Normality

- Multivariate normality – Each variable is normally distributed and all linear combinations of the variables as normally distributed.
  - Not usually tested
  - Tests that are used tend to be overly sensitive.
  - Residuals



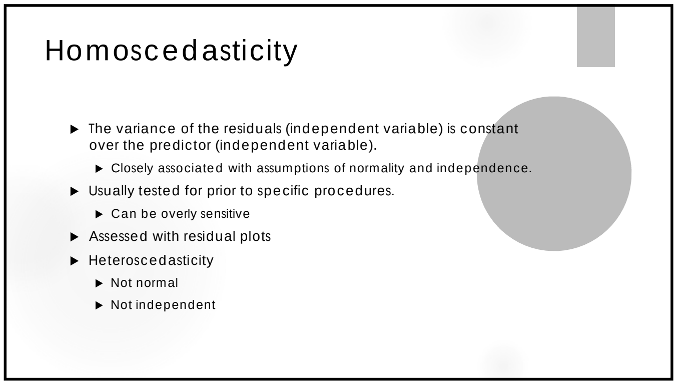
Normality

- Normality is assessed
  - Tests of skew and kurtosis
  - Visual assessment
  - Normality and independence of residuals



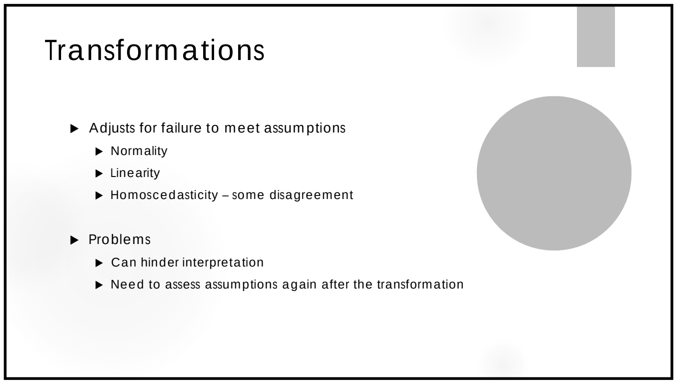
Linearity

- The coefficients of the variate are constant – straight line relationships.
- Detection
  - Residual plots
    - Residuals on the y axis and predicted values on the x axis
    - Biplots



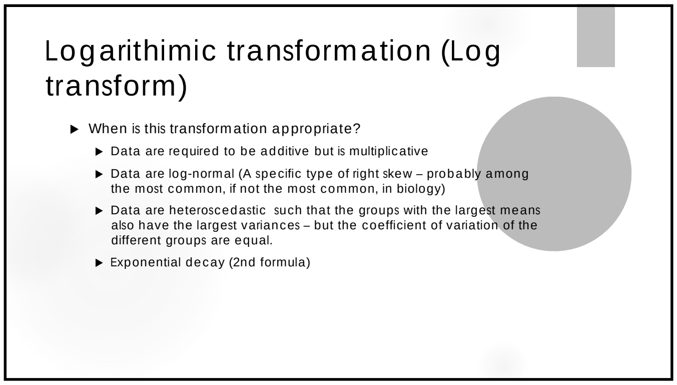
Homoscedasticity

- The variance of the residuals (independent variable) is constant over the predictor (independent variable).
  - Closely associated with assumptions of normality and independence.
- Usually tested for prior to specific procedures.
  - Can be overly sensitive
- Assessed with residual plots
- Heteroscedasticity
  - Not normal
  - Not independent



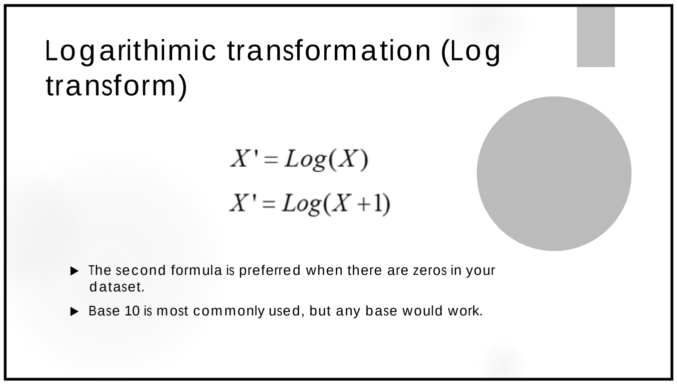
Transformations

- Adjusts for failure to meet assumptions
  - Normality
  - Linearity
  - Homoscedasticity – some disagreement
- Problems
  - Can hinder interpretation
  - Need to assess assumptions again after the transformation



Logarithimic transformation (Log transform)

- When is this transformation appropriate?
  - Data are required to be additive but is multiplicative
  - Data are log-normal (A specific type of right skew – probably among the most common, if not the most common, in biology)
  - Data are heteroscedastic such that the groups with the largest means also have the largest variances – but the coefficient of variation of the different groups are equal.
  - Exponential decay (2nd formula)

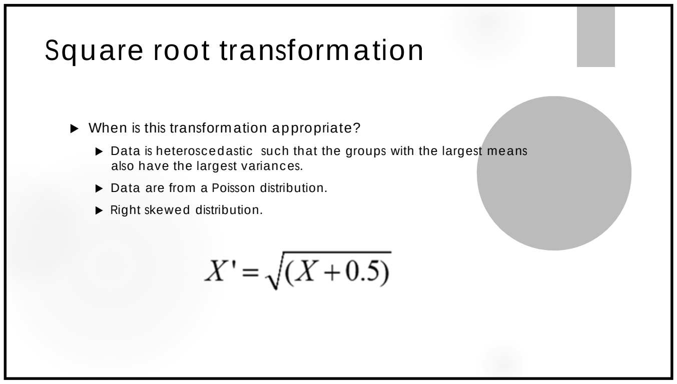


Logarithimic transformation (Log transform)

$$X' = Log(X)$$

$$X' = Log(X+1)$$

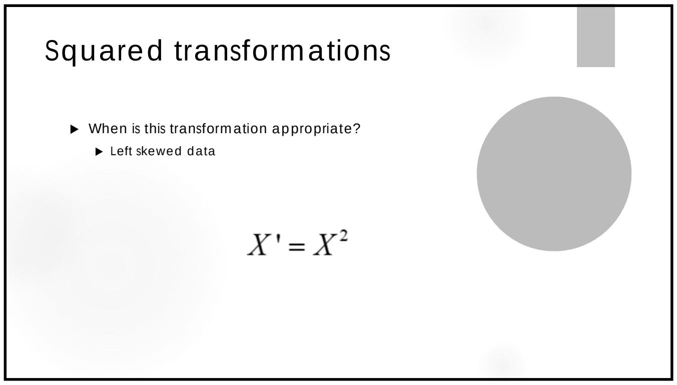
- The second formula is preferred when there are zeros in your dataset.
- Base 10 is most commonly used, but any base would work.



Square root transformation

- When is this transformation appropriate?
  - Data is heteroscedastic such that the groups with the largest means also have the largest variances.
  - Data are from a Poisson distribution.
  - Right skewed distribution.

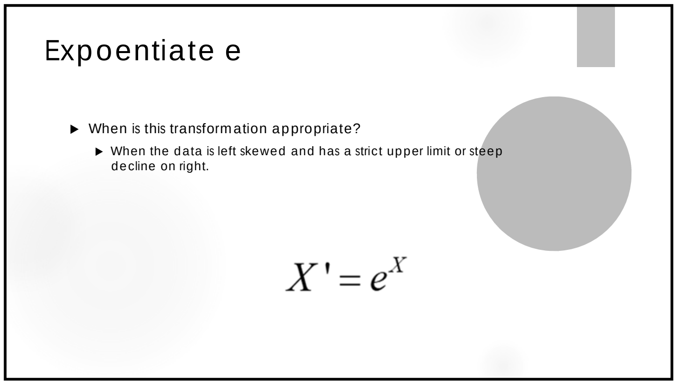
$$X' = \sqrt{(X+0.5)}$$



Squared transformations

- When is this transformation appropriate?
  - Left skewed data

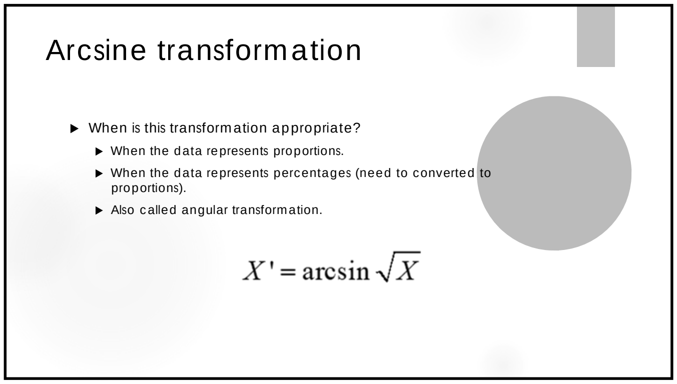
$$X' = X^2$$



Expoentiate e

- When is this transformation appropriate?
  - When the data is left skewed and has a strict upper limit or steep decline on right.

$$X' = e^X$$



Arcsine transformation

- When is this transformation appropriate?
  - When the data represents proportions.
  - When the data represents percentages (need to converted to proportions).
  - Also called angular transformation.

$$X' = \arcsin\sqrt{X}$$

Transformation

- Reflection
  - Preferred by some when you have a left skew.
  - Opposite interpretation

$$X' = Max(X) + 1 - X_i$$

Multicollinearity

- Variables that are highly correlated – typically independent or predictor variables.
- Identifying variables with multicollinearity
  - Correlation matrices (r> 0.90)
- Dealing with multicollinearity
  - Drop all but one of the highly correlated variables
  - Unless doing factor analysis (PCA), you should not include redundant variables

Singularity

- A variable is the same as another variable or a linear combination of several other variables.
- Handled like multicollinearity.