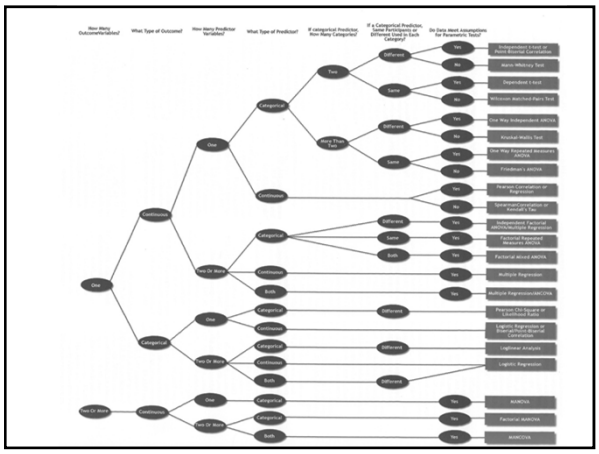
Overview of the General Linear Model

Goals

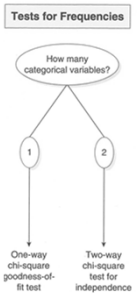
- Selecting a statistical test
- Relationships among major statistical methods
- General Linear Model and multiple regression
- Special cases of multiple regression

Questions to consider when selecting a statistical test

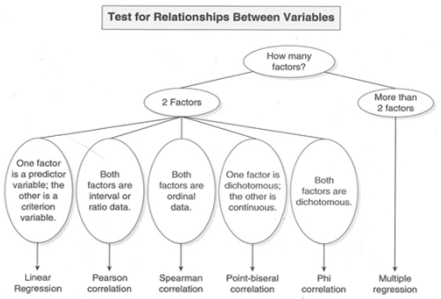
- Major research question?
- How many outcome variables?
- What type of outcome variables?
- How many predictor variables?
- What type of predictor variables?
- If a categorical predictor, how many categories?
- If a categorical predictor, same or different participants used in each category?
- Inclusion of covariates?
- Do the data meet assumptions of parametric tests?



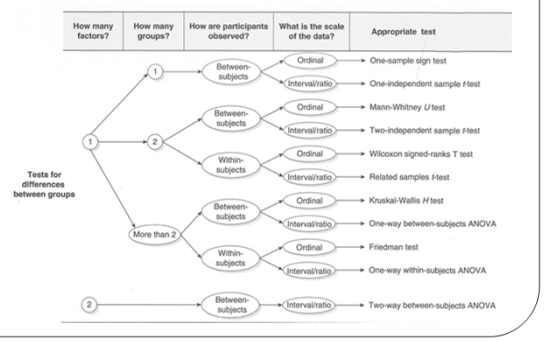
Major Research Question: Frequency of Occurrence



Major Research Question: Degree of Relationship Among Variables



Major Research Question: Differences Between Groups



Relationships Among Major Statistical Methods

- About 90% of psychology articles focus on t-tests, ANOVAs, correlations, or regressions
- We have focused on the differences between these types of tests…but they are actually very similar
  - Their similarity is due to the fact that they are all derived from the same general formula which is known as the **General Linear Model (GLM)**

Relationships Among Major Statistical Methods

- The most general of these tests is multiple regression
- The other three tests are simply special cases of multiple regression
  - "Special case" means that the formula for the more specific tests (e.g., ANOVA) can be derived from the formula for multiple regression
  - You will get the same basic results if you use the more specialized or more general test
  - If you could only use one statistical test for the rest of your career, then you would probably want to choose multiple regression because it is the most versatile

The General Linear Model (GLM)

- The GLM has two primary assumptions
  - Linearity: Pairs of variables are assumed to have a linear relationship with each other (i.e., can be represented by a straight line)
    - The GLM can also deal with curvilinear and multiplicative relationships (as well as other types of data such as categorical variables)
  - Additivity: Additional predictor variables are assumed to add predictability to earlier predictor variables
    - Multivariate models consist of weighted terms (predictor variables) that are added together

Requirements of the General Linear Model

- There must be a set of "participants," "cases," or "units"
- Each of these participants will have values or measurements on two or more variables
  - Data can be numerical/continuous, dichotomous, or multicategorical
  - Participants will usually serve as "rows" and variables will serve as "columns"
- Each variable should be able to be represented by a single column of numerical values
  - We will talk about coding strategies later (e.g., dummy coding)
- Each analysis should have one outcome variable
  - …but it can have more than one predictor variable or multiple covariates
- The outcome variable must be numerical/continuous

Flexibility of the General Linear Model

- A variable may be a natural property of a participant (e.g., age) or a property that is manipulated in an experiment (e.g., experimental condition vs. control condition)
  - Manipulated variables are usually categorical but can be numerical/continuous (e.g., number of acts of violence a participant is exposed to during a study)
- Researchers can conduct multiple analyses using the same set of variables
- GLM does not distinguish between predictor variables and covariates
- Predictor variables and covariates may be numerical/continuous, dichotomous, or multicategorical
- Predictor variables and covariates may be intercorrelated
- Predictor variables and covariates may interact when predicting the outcome variable
- Despite being a "linear" model, it can easily be extended to situations that involve curvilinear associations between variables

The General Linear Model (GLM)

- The GLM is based on prediction (i.e., regression)
- A regression equation represents the value of a criterion variable (Y) as a combination of one or more predictor variables (Xs) plus error
  - Simplest form is bivariate regression
  - $Y = BX + A + e$
    - B is slope (the change in Y associated with a one-unit change in X)
      - B is also known as the "unstandardized regression coefficient" or the "unstandardized regression weight"
    - A is the Y-intercept (a constant representing the value of Y when X is 0)
    - e is a random variable representing error of prediction

Simple Bivariate Form of GLM

- If X and Y are converted to z-scores, then the bivariate regression simplifies to the following
  - $z_Y = \beta z_X + e$
    - Y-intercept term drops out because the line crosses y-axis at 0 because of the standardization
    - β is the standardized slope (or standardized regression coefficient) which represents the strength of the relationship between X and Y
      - In bivariate regression, β is equal to the Pearson product-moment correlation coefficient (the closer β is to +1 or -1, then the better the prediction of Y from X [or X from Y])

Simple Bivariate Form of GLM

- An important issue for the selection of statistical tests is whether the data are continuous or discrete
- Three forms of bivariate regression
  - X continuous, Y continuous: **Pearson product-moment correlation**
  - X dichotomous, Y continuous: **Point biserial correlation**
  - X dichotomous, Y dichotomous: **Phi coefficient** (related to chi-square)
    - If dichotomous variable is coded as 0 and 1, then these are all identical

Simple Multivariate Form of GLM

- The first generalization of the simple bivariate GLM is to increase the number of predictor variables
  - Additivity is important for understanding this extension of the GLM
  - $z_Y = \sum_{i=1}^{k} \beta_i z_{X_i} + e$
    - This formula is telling us to use the weighted sum of the Xs
      - Example of three predictors:
        - $z_Y = \beta_1 z_{X_1} + \beta_2 z_{X_2} + \beta_3 z_{X_3} + e$
    - If Y is continuous and all of the Xs are continuous, then this is multiple regression
    - If Y is continuous and all of the Xs are discrete, then this is the special case of ANOVA
      - The values of X represent group membership and the emphasis is on finding mean differences in Y (rather than predicting Y)…but the basic equation is the same because a significant difference among groups implies that information about X can be used to predict performance on Y

Simple Multivariate Form of GLM

- All Xs continuous, Y continuous: **Multiple regression**
- All Xs discrete, Y continuous: **ANOVA**
- Some Xs continuous and some discrete, Y continuous: **ANCOVA**
- All Xs continuous, Y dichotomous: **Two-group discriminant analysis**
- All Xs discrete, Y is category frequency: **Multiway frequency analysis**
- Xs continuous or discrete, Y dichotomous: **Two-group logistic regression**
- Xs at each level may be continuous or discrete, Ys at each level are continuous: **Multilevel modeling**
- Xs continuous and/or dichotomous, Y continuous (time): **Survival analysis**
- Xs continuous (time) and dichotomous, Y continuous: **Time series analysis**

Full Multivariate Form of GLM

- GLM can also deal with situations where there are more than one outcome variables (multiple Ys)
  - We will revisit this issue later in the semester when everyone is more familiar with regression

Full Multivariate Form of GLM

- All Xs continuous, Ys continuous: **Canonical correlation**
- All Xs discrete, Ys continuous: **MANOVA**
- Some Xs continuous and some discrete, Ys continuous: **MANCOVA**
- All Xs discrete, all Ys continuous and commensurate (i.e., measured on the same scale): **Profile analysis**
- All Xs continuous, all Ys are discrete: **Discriminant analysis**
- All Xs latent, all Ys continuous: **Factor analysis /principal components analysis**
- Xs continuous and/or latent, Ys continuous and/or latent: **Structural equations modeling**
- All Xs discrete, Y is category frequency: **Multiway frequency analysis**
- Xs continuous and/or discrete, Y discrete: **Polychotomous logistic regression analysis**