Overview of Statistics 572

Bret Larget

Departments of Botany and of Statistics
University of Wisconsin—Madison

January 23, 2007

Welcome to Statistics 572!

- Introduce self and TAs.
  - Bret Larget
  - Heather Brazeau
  - Yali Wang
- Comment on syllabus.
  - Textbook
  - Web for notes and grades (print notes before lecture)
  - Objectives
  - Computing (go R!)
  - Assignments (late policy)
  - Exams (save dates)
  - Grading
  - Academic honesty
  - Discussion sections (attend the one you want)

The Big Picture

- A statistical approach to data analysis can lend *insight to biological understanding* of a wide variety of problems.
- In a statistical approach, measurable variables are treated *as realizations from a model* that relates biological meaningful parameters and stochastic sources of variation.
- No model accounts for all aspects of the underlying biology, but *an appropriately selected model can be very useful*.
- Many data analysis problems arising from the biological sciences *are appropriate for linear and generalized linear models*, a rich family of possible models.

Variables

- Typically, one variable of interest is modeled as a *response variable* which is related to one or more *explanatory variables*.
- Variables can be categorized as *quantitative* or *categorical*.
- Quantitative variables are typically either measured on a *continuous scale* or are *discrete*, variables that are *counts*.
- The appropriate choice of model is determined in part by the *types of the response and explanatory variables*.
- A *linear combination* of the variables $X_1, \dots, X_k$ takes the form

$$\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

- Linear and generalized linear models include *linear combinations of explanatory variables*.

Examples of Linear Models

- *Simple Linear Regression.—*

  response variable: continuous quantitative variable
  explanatory variable: one quantitative variable
  error structure: normal distribution
  model: $y_i = \beta_0 + \beta_1 x_i + e_i, \qquad e_i \sim \mathrm{N}(0, \sigma^2)$
  example: response variable is phosphorous concentration in plant tissue, explanatory variable is phosphorous concentration in the soil.

- *Multiple Linear Regression.—*

  response variable: continuous quantitative variable
  explanatory variables: more than one quantitative variables
  error structure: normal distribution
  model: $y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + e_i, \qquad e_i \sim \mathrm{N}(0, \sigma^2)$
  example: response variable is soybean yield, explanatory variables are hours of daylight and amount of nitrogen.

Examples of Linear Models (cont.)

- *One-way ANOVA.—*

  response variable: continuous quantitative variable
  explanatory variable: one categorical variable
  error structure: normal distribution
  model: $y_{ij} = \alpha_i + e_{ij}, \qquad e_{ij} \sim \mathrm{N}(0, \sigma^2)$
  example: response variable is milk yield explanatory variable is diet (four treatments)

- *Multi-way ANOVA.—*

  response variable: continuous quantitative variable
  explanatory variables: more than one categorical variables
  error structure: normal distribution
  model: $y_{ijk} = \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijk}, \qquad e_{ijk} \sim \mathrm{N}(0, \sigma^2)$
  example: response variable nitrogen level in manure, explanatory variables are diet treatment, period, and interaction.

Examples of Linear Models (cont.)

- *Linear models with both types.—*

  response variable: continuous quantitative variable
  explanatory variables: both quantitative and categorical
  error structure: normal distribution
  model: $y_{ij} = \beta_0 + \beta_1 x_{ij} + \alpha_i + e_{ij}, \qquad e_{ij} \sim \mathrm{N}(0, \sigma^2)$
  example: response variable is milk yield, explanatory variables are diet (four treatments) and days in milk.

- *Polynomial regression.—*

  response variable: continuous quantitative variable
  explanatory variables: single quantitative explanatory variable
  error structure: normal distribution
  model: $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + e_i, e_i \sim \mathrm{N}(0, \sigma^2)$
  example: response variable is disease area, explanatory variable is age

Examples of Linear Models (cont.)

- *Mixed models.—*

  response variable: continuous quantitative variable
  explanatory variables: variables of both *fixed* and *random* effect.
  error structure: normal distribution
  model: $y_{ij} = \beta_0 + \beta_1 x_{ij} + a_i + e_{ij}, e_{ij} \sim \mathrm{N}(0, \sigma^2), a_i \sim \mathrm{N}(0, \sigma_a^2)$
  example: response variable is percentage cover of vegetation, site is modeled as a random effect, quantitative variables include soil moisture.

- *Repeated measures.—*

  response variable: continuous quantitative variable
  explanatory variables: one or more including random effect for individual
  error structure: normal distribution
  example: response variable is hormone concentration, explanatory variables include individual and day.

Examples of Generalized Linear Models

- *Logistic Regression.—*

  response variable: categorical variable with two levels
  explanatory variables: one or more
  error structure: binomial
  model: $\mathrm{P}\{y_i = 1\}$ is a function of $\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$.
  example: response variable is seed germination, explanatory variables include temperature and treatment.

- *Poisson regression.—*

  response variable: non-negative integer-valued variable
  explanatory variables: one or more
  error structure: Poisson
  model: $\mathrm{P}\{y_i = k\}$ is a function of $\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$.
  example: response variable is number of seeds produced, explanatory variables include treatment and light intensity.

Data request

- I will present each type of model with an example and data.
- These case studies will be more interesting if they are related to *genuine research problems*.
- If you or someone in your lab *has data that falls into the scope of these models*, and you are willing/able to share, *please contact me*.