☼ RNR 613 — Introduction and Background

"There are three kinds of lies: lies, damned lies, and statistics" — *Benjamin Disraeli*

What is (the study of) *Statistics*? The study of methods to collect and interpret scientific information using probability to address the uncertainty of this information.

Why use Statistics?

Data Collection

• Collect data with minimum bias and maximum precision. ➛ *Sampling Methods*
• Design experiments in such a way as to maximize the chances of detecting biologically important effects (i.e., with high statistical power = low Type II error rate) while controlling the chances of drawing incorrect conclusions (low Type I error rate). ➛ *Experimental Design*

Data Analysis

• Summarize scientific information. ➛ *Descriptive Statistics*
• Estimate population parameters using sample data. ➛ *Parameter Estimation*
• Test hypotheses. ➛ *Inferential Statistics*

Definitions

• *Population* — The entire collection of entities about which one wishes to make an *inference* or draw a *conclusion* about (also called aggregate or universe).
• *Sample* — A subset of a population. Used because we usually cannot measure all individuals in a population. It is the *sample* we observe, but the *population* we wish to know.
• *Simple Random Sample* — A sample of size *n* from a larger population selected in such a way that every sample of size *n* has the same chance of being selected.
• *Parameter* — The *true* value of some population attribute, which is almost always unknown; or an unknown *constant* that describes a key feature in a model for answering a question of interest. Parameters are often represented by Greek letters, such as μ for the population mean, and σ for the population standard deviation.
• *Statistic* — Any quantity that is computed or estimated from sample observations. Statistics or estimates are represented by Roman letters, such as $\bar{Y}$ for sample mean and *s* for sample standard deviation; statistics are sometimes distinguished from the parameter they estimate by a "hat," such as $\hat{\sigma}$.
• *Probability* — Set of mathematical tools to quantify concepts we understand intuitively, such as "likelihood", "predictability",and "certainty." We use probability to gauge the amount of confidence to place on sample estimates.
• *Model* — Some approximation of reality.
• *Statistical model* — A mathematical expression that help us predict a *response variable* as a function of one or more *explanatory variables*, based on a set of assumptions. These assumptions allow the model not to fit exactly, and are made about random terms in the model called *error* (ϵ).

E.g., use sample data to develop a statistical model to predict how a response variable (say, nestling mass) varies with changes in an explanatory variable (say, nestling age): $y = b_0 + b_1x + \epsilon$, where *y* represents nestling mass, b_0 and b_1 are estimates of model parameters for intercept (β_0) and slope (β_1) that describe the line fit using sample data; ϵ represents random error = that portion of observed data that the model does not describe exactly.

Assessing Model Appropriateness and Model Fit

Strive for "acceptable" model fit; examine assumptions of a particular model with the data collected to determine if they are met reasonably. We will choose models based on the *context* (data type and sampling or experimental design) then use graphical tools to determine the chosen model's appropriateness.

☼ RNR 613 — Variables, Data Tables

Statistics — Study of methods to collect and interpret scientific information using probability as the foundation by which to address the uncertainty of this information.

• Types of Variables (or in JMP-speak, modeling type)

Continuous—Data values for a variable are measured on a continuous scale. E.g., *body mass* is often measured and recorded on a *continuous* scale, where values such as 40 g or 3,154.2 g are acceptable. It can be useful to distinguish between *continuous* and *discrete* data. *Continuous* data can be represented with any and all conceivable values within a particular range, such as the height of a plant being 36.354 cm; *discrete* data (or *meristic* data) can be represented by only certain values within a particular range, such as *number of leaves on a plant*, where 22, 185, or 45 are possible, but 22.8 is not. Count data are discrete but sometimes can be modeled as continuous.

Ordinal—Data values for a variable are *labels* identifying a category and their *order is meaningful*. E.g., a person's *highest educational level* might be recorded ordinal, where the categories of interest might be grade school, high school, college, and graduate school. Another way to view these data is as *rank ordered*, where the actual values for the variables are compressed into meaningful (but probably somewhat arbitrary) categories. What often is recorded as *ordinal* data may have been recorded as *continuous* data if measurements were made more precisely. E.g., we could replace *education level* above with the actual *number of years a person spent in school*, and then model the variable as *continuous* (and *discrete*).

Nominal—When data values for a variable are *labels* identifying a category and their *order is not meaningful*. E.g., attributes such as *color* are nominal (meaning named) because the categories do not represent some underlying, quantitative scale. Although *colors* such as blue, red, and yellow are different (which is usually what we are interest in), they are not *quantitatively* different from each other (i.e., order of categories has no particular meaning).

As mentioned above, It is sometimes possible to consider certain variables as more than one type. For example, if *age* is measured in years, you might consider *age* to be either *continuous* or *ordinal*. In these cases there is not a right or wrong way to model the variable; however, there may be certain advantages to one approach over another. What is critical, however, is to *note when there is order in data* that are categorical in nature. Considering something *nominal* when it is really *ordinal* can result in a less efficient analysis or result in limiting yourself to fewer modeling options.

• Creating Data Tables that Facilitate Analysis

One obstacle to efficient data analysis exists when data not organized carefully within a statistical software's spreadsheet. By considering carefully the best way in which to enter data, analyses are facilitated. Almost all software packages handle a dataset as a matrix of rows and columns. Typically, each observation or sample becomes a row in a data table and each attribute or measurement associated with that observation becomes a column. E.g., you are studying 5 attributes of 3 different plant species from 2 geographic regions. You have taken measurements of all attributes from 10 individuals of each species in each region. Your data table should have 7 columns (species, region, attrib1-attrib5) and 60 rows (3 species x 2 regions x 10 individuals = 60).

There are plenty of circumstances when this approach will not work, most commonly when you measure the same individual (sampling or experimental unit) more than once, which you might do for at least 2 reasons: (1) sampling units are highly variable, so you must take multiple measurements to reduce sampling variability (subsampling), or (2) your design call for repeatedly measuring the same units through time (repeated measures).

You then have 2 alternatives. First, you can add a column to identify uniquely each sample/ experimental unit and include a complete set of measurements (rows) for each subsample. E.g., you have 20 field plots where you measured a set of attributes, but because the plots are large, so you take 3 subsamples per plot. Your dataset would then have 20 plots x 3 subplots/plot = 60 rows. Second, you could create additional columns for all measurement within a sample. So if you measured 3 attributes in the above scenario with 3 subsamples per plot, you would have 3 attributes x 3 subsamples, or 9 columns and 20 rows.