Statistics Review

Consider the general question, **"What justifies a conclusion?"**

In the sciences, the answer to this question must involve the notion of **Valid Methodology**. Valid inference must rest on valid methodology. There is a tight link between the population of interest, how it is sampled and the (inductive) inferences being made from the sample data to the population sampled. Sampling design, data collection and proper and objective data analysis must receive careful attention to justify a result or conclusion or "statistical inference."

Inductive inference is going "from the particular to the general" or, in the case of sampling from a population, "inference from the sample to the population." A theorem in logic states that there is always uncertainty in inductive inference. There are no <u>certain</u> generalizations. Such uncertainty can be measured by statistical methods if the sample have been properly drawn and assumptions of the method have been met.

Measures of precision are essential in inductive inference. Precision relates to repeatability. If an estimator is very precise, one expects a high degree of repeatability in estimates from still other samples. An estimate $\hat{\theta}$ (the ∧ denotes an estimate or estimator) of a parameter θ without a measure of its precision is nearly useless. Typical (frequentist) measures of precision include:

Sampling variance, $\text{var}(\hat{\theta})$

Standard error, $\text{se}(\hat{\theta})$

Coefficient of variation, $\text{cv}(\hat{\theta}) = \hat{\text{se}}(\hat{\theta})/\hat{\theta}$ (often seen expressed as a percent)

Confidence interval (often asymmetric)

Often the estimates of sampling variance or standard error are *conditional* on a given model. Then, this can be noted more exactly as $\text{var}(\hat{\theta} \mid M_i)$ to denote the variance, given model M_i.

Parameters are often denoted generically by θ. Then θ can be a conditional, finite survival probability (S), on a population size (N), or the mean life span ($\bar{L}$), mean reporting probability (μ_r), or the standard deviation of survival parameters (σ_s), or the rate of finite population increase (λ). Of course, one must define such parameters to allow the reader to understand the notation being used.

Why Theory?

It is foolish and inefficient to proceed without a formal theory! The Breeding Bird Survey (BBS) and the Dove Call Count Survey are examples where theory was not carefully considered before field methods were implemented. Sampling design and field protocol were seriously flawed and no rigorous analysis theory was considered (e.g., in the case of the Breeding Bird Survey, a semi-random sample of roads was used; thus inference must be restricted to "road-side habitats"). The price for these poor practices has been very high. This is a "cart before the horse" problem.

Theory provides a basis for planning a survey, selecting a sampling design, refining a field protocol. Theory provides a means for deriving estimators with optimal properties (e.g., survival and recapture probabilities, population size, finite rate of population growth). Theory is needed for deriving estimators of precision (e.g., the variance-covariance matrix). Theory is particularly important in selecting a "best" approximating model (or models) for inference about a population parameter. <u>Model</u> <u>selection</u> <u>is</u> <u>a</u> <u>key</u> <u>component</u> <u>in</u> <u>the</u> <u>analysis</u> <u>of</u> <u>empirical</u> <u>data.</u>

Data analysis methods should be objective and repeatable. *Ad hoc* methods are not reliable and, therefore, not very useful in the sciences. Wildlife and fisheries has made frequent use of "index" values and this represents poor practice because the index is a function of factors such as observer variability, the environment, and types of cues from the species being surveyed. Index values should be avoided in serious work.

Theory must always be based on some assumptions. The assumptions must be known and field protocol must assure that assumptions are met as well as possible. Without theory, one often does not know what assumptions are the basis of the method being considered. Thus, field protocols ignore these issues and analysis becomes a salvage job of "the numbers."

A critical function of statistical science is to provide a valid formalism for making inductive inference and estimating the associated uncertainty. Statistical science is not so much a branch of mathematics as it is an area of science concerned with the development of a practical theory of information.

Some Important Definitions:

Population The individuals of interest, the group of animals to which inferences are to be made. In other sampling studies, the population might be cars, banks, vegetation types, etc. The investigator defines the population of interest (e.g., the woodpeckers in Larimer County, CO in July, 1992). Unless the population is well defined, sampling may be done in such a way that causes inferences to be weak or even incorrect.

Sample Data available on some fraction of the defined population of interest. Often, data are collected on only a very small proportion of the population (e.g., perhaps < 1%).

Typically, when sampling biological populations, the sampling fraction is not known and must be estimated.

Census A complete enumeration of all individuals in a defined population. This can rarely be done with biological populations. Note, birders often misuse this word when they actually mean "survey" or "sample." Only rarely is a census of a biological population at all feasible. Census data eliminate the concept of repeatability and its measures such as sampling variances (repeatability with a census is perfect, by definition, and the sampling variance is zero as no sampling was needed).

Purpose of sampling Most commonly, an investigator wants to estimate some parameter associated with a population of interest (e.g., the proportion of the population of individuals that is male or the marked population's mean life span).

Statistical Inference An inference made from the sample data to the defined population using inductive methods and statistical theory. For example, if 700 to 1,100 red-cockaded woodpeckers are banded annually for 12 years and valid methods are used to make estimates of the average annual probability of survival; then one can made a statistical inference about the banded population. If one wants to assume that the banded population represents the unbanded population; then an inference can be made about the survival probability of the population in general for the years to study, but this is not usually a formal statistical inference. Similarly, if one concluded something about the survival probability of the woodpeckers for years other than the study period, this would not constitute a statistical inference.

Assumptions In empirical problems, one must make certain assumptions (e.g., independence of marked individuals). A goal of the statistician is to avoid methods that are very sensitive to the partial failure of a particular assumption (methods that tend to be insensitive to the breakdown of certain assumptions are called "robust").