

## Lec 02: Estimation & Hypothesis Testing in Animal Ecology

### Parameter Estimation from Samples

#### Samples

We typically observe systems incompletely, i.e., we sample according to a designed protocol. We then apply mathematical formulas to the data to obtain parameter estimates for quantities of interest.

Because only a sample is collected, estimates obtained are subject to *sampling variability*. For example, if we study annual survival rate for 10 animals from a population that has a true annual survival rate of 0.5, we might find that only 3 animals survive even though if we knew the true survival rate we would expect 5 of 10 to survive. Such a departure is due to *sampling variability*.

1. Define the statistical population as the collection of individuals that can be potentially sampled.
2. Apply a sampling design to the population whereby probabilities can be ascribed to samples.
3. Specify characteristics to be measured on sampled individuals.

#### Distributions

We then think of natural variation in the organisms in terms of some underlying pattern that follow some *frequency distribution* (how common are the different values that can be observed). That frequency distribution reflects a *probability distribution* when individuals are sampled randomly.

Distributions can be broken into *discrete distributions* and *continuous distributions*. Within each of these types, we can further identify various distributions.

Some classic examples of discrete distributions include the *binomial distribution* (random events for which one of two outcomes can occur, e.g., live or die), or the *multinomial distribution* (random events for which >2 outcomes can occur, e.g., outcome of the the roll of a 6-sided die). Those two distributions will be used extensively in this course and you will become quite familiar with them.

Many data in biological samples are continuous in nature. For continuous data the probability distributions are smooth distribution functions over a range of values appropriate for the data. The most heavily used continuous distribution is the univariate normal distribution with probability density function

$$f(x | \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \cdot \exp\left[-\frac{1}{2} \cdot \left(\frac{x - \mu}{\sigma}\right)^2\right]$$

The normal distribution is parameterized by its mean ( $\mu$ ) and variance ( $\sigma^2$ ). As you should know it is symmetric about its mean, is bell-shaped, and is more or less peaked depending on the variance.

#### Distribution Parameters

The *expected value* of a random variable  $x$  is the average of values that  $x$  can take, with each value weighted by the frequency of occurrence of the value:  $E(x) = \sum_x x \cdot f(x | \theta)$

For the roll of a die  $E(x) = 1(1/6) + 2(1/6) + \dots + 6(1/6) = 3.5$

Parameters of a distribution such as the mean or variance are typically not known but have to be estimated from sample data collected from the population. Thus, we do things like estimate the population mean based on the sample mean.

***The theory of statistical inference deals with sample-based inferences about the parameters of a statistical population and the degree of confidence with which inferences can be made.***

#### *Replication and Statistical Independence*

Replicate samples help us to better characterize a population and to assess the variability in our sampling procedure (how repeatable is it).

Statistical independence occurs if the value of one random variable tells us nothing about the value of another beyond the fact that they were generated from the same random process.

When various random variables are *not* independent, they are correlated or covary, i.e., take values that are associated (either positively or negatively). Although most graduate students are familiar with correlation, many beginning students are less familiar with covariance:

$$\text{cov}(x, y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

Covariance, which is not bounded between -1 and 1, relates to correlation as follows:

$$\text{corr}(x_1, x_2) = \frac{\text{cov}(x_1, x_2)}{SD(x_1)SD(x_2)}$$

Covariance is typically more informative if it is referenced to the underlying variation in each variable. Thus, one useful way of describing variables is via a *variance-covariance matrix* (sometimes called  $\Sigma$ ).

For example,  $\Sigma = \begin{bmatrix} 0.041 & 0.004 & -0.005 \\ 0.004 & 0.044 & -0.013 \\ -0.005 & -0.013 & 0.049 \end{bmatrix}$

The correlation matrix for this variance-covariance matrix is:

$$\text{corr} = \begin{bmatrix} 1.000 & 0.105 & -0.123 \\ 0.105 & 1.000 & -0.277 \\ -0.123 & -0.277 & 1.000 \end{bmatrix}$$

### **Parameter Estimation**

#### *Two Questions*

1. The probability question – given a probability distribution with known parameters, how likely are various outcomes?
2. The estimation question – given observed data, what is the distribution from which the data arose? And, in our problems in this course, we'll usually ask, "given observed data, what is (are) the corresponding value(s) of  $\theta$  that parameterize the distribution that gave rise to those data."

## Bias, Precision, and Accuracy

If replicate samples were obtained, our estimates would vary and any given estimate will differ from other estimates and from the 'true' value of the parameter of interest. Thus, the question arises: how good is a given estimator in the sense of being close to the true parameter value. We measure this by considering estimator bias, precision, and accuracy.

### Estimator Bias

If  $E(\hat{\theta}) \neq \theta$ , then the estimator is biased with bias equal to the difference between  $E(\hat{\theta})$  and  $\theta$ . Note, however, that this is a feature not of one estimate resulting from the estimator but the long-run behavior of the estimator.

It turns out that some estimators are, in fact, biased. Some estimators become quite biased if some of the assumptions underlying the estimator are violated, i.e., they are not *robust* to violations of some or all assumptions.

### Precision

The tendency of replicated estimates to be dispersed is an expression of estimator precision, which is measured by the variance of the estimator.

$$\text{var}(\hat{\theta}) = E\{[\hat{\theta} - E(\hat{\theta})]^2\}$$

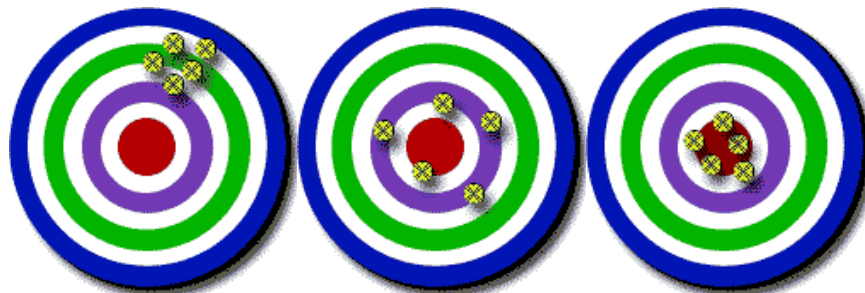
If replicate estimates are widely dispersed, the estimator has low precision. It turns out that in most cases, we don't have replicated datasets on which to estimate *estimator precision*. Rather, we use other statistical methods to obtain estimates of variance.

### Accuracy

Precision and bias do not have to be related, e.g., you can have a highly precise, biased estimate. Accuracy combines both bias and precision as an assessment of estimator performance. One method of doing so is mean squared error (MSE), which is defined as:

$$\text{MSE}(\hat{\theta}) = \text{var}(\hat{\theta}) + (\text{bias}(\hat{\theta}))^2$$

Thus, an accurate estimator is precise and unbiased. An inaccurate estimator can be biased, imprecise, or both. Also, it is common in the statistics literature to refer to RMSE or the root mean squared error or the square root of MSE. Computer simulation or replicate samples are approaches used to evaluate estimator performance.



## Estimation Procedures

We will use *Maximum Likelihood Estimation* throughout this course. It will be presented by using examples from the binomial distribution and later through the multinomial distribution. One presumes to know the mathematical form of the distribution function  $f(x|\theta)$ , but not to know the actual value of  $\theta$  for the distribution. We will express the likelihood function as  $L(\theta|x)$ , which emphasizes that we are trying to find the most likely parameter estimates given an observed dataset.

### **Confidence Intervals and Interval Estimation**

How confident are you that the estimate obtained from random sampling accurately represents the actual parameter? We will address the question using confidence intervals for estimates. In a formal sense, we will seek to construct confidence intervals such that in replicated studies, confidence intervals about estimated parameters will include the true parameter value a stated percentage of the time. It is common to construct 95% confidence intervals: in such a case, if a study were repeated 100 times and a 95% CI were constructed each time, we would expect 95 of the 100 confidence intervals to contain the actual parameter value.

There are a variety of ways of constructing confidence intervals. One common method that we will encounter is the construction of asymptotically normal CIs

$$\hat{\theta} \pm 1.96 \cdot SE(\hat{\theta})$$

The method is based on the MLE for the parameter being within 1.96 standard deviations of the parameter with probability 0.95. This approach relies on asymptotic properties of MLEs, but factors such as small sample sizes may cause CIs created this way to have coverage levels below the stated level. An alternative is to work with the actual likelihood profile to create CIs. CIs from the profile likelihood approach are often asymmetric about the estimated parameter value.

### **Hypothesis Testing**

#### *Significance Testing*

Statistical testing of hypotheses is closely associated with CI estimation. Introductory statistics classes for graduate students typically cover the topic of hypothesis testing in terms of comparing a null and alternative hypothesis. The topics of Type I errors (rejecting a true null), Type II errors (failing to reject a false null), and statistical power (i.e., the ability of a test to reject a false null hypothesis) are all germane to such work: these are typically covered quite well in introductory statistics classes for graduate students as well. These topics are important to understand but will not be covered in this course as our focus will be on evaluating multiple models with information-theoretic approaches.

#### *Goodness-of-Fit Tests*

We would often like to be able to determine how adequate a statistical model is at characterizing our field data. For example, does a model that estimates a single annual survival rate adequate for describing survival obtained from male and female animals of varying ages from 5 different years? Here, we are evaluating a null hypothesis that the particular model being used fits the data. Rejection of the null hypothesis occurs if significant lack of fit is in evidence. We will use several goodness-of-fit (GOF)

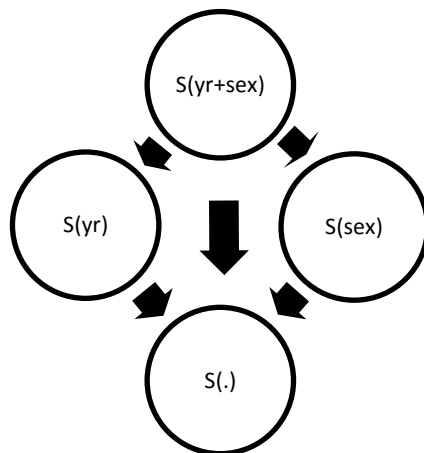
procedures in the course, learn of why one cares about GOF in model selection problems, and discuss what might be done in cases where lack of fit appears to be a problem.

### *Likelihood Ratio Tests for Model Comparisons*

We will be interested in comparing >2 candidate models as well as using GOF tests to assess the adequacy of models to characterize the data. It is possible to employ model comparisons that can be seen as a type of hypothesis test: we can use a test that compares the fit of a simpler model against the fit of an alternative and more general (more complex) model. Here, we would test null hypothesis that the simpler model fits the data as well as the more complex model. It is common to proceed as follows:

1. Conduct GOF on a general model
2. Calculate likelihoods for various versions of the general model, i.e., models that constrain parameters to greater or lesser extents. For example, one could calculate the likelihood for a model that allows annual survival rate to differ by year, sex, and age class. One could then also calculate the likelihood for (a) a model that constrains survival rate to be constant across years but to vary by sex and age class or (b) a model that constrains survival rate to be constant for all animals in all years.
3. Compare the likelihoods using likelihood ratio tests and evaluate the statistical significance based on the difference in the number of parameters in the two models being compared.

It is noteworthy that the models must be nested versions of one another for this procedure to work.



Here, all models are simplifications of the models above and that any such comparisons can be made with likelihood ratio tests. But note, that one can't compare models that are across from one another in the drawing. That is, a model ( $S_{yr}$ ) that allows to survival rate to vary by year but constrains male and female survival rates to be the same cannot be compared to one ( $S_{sex}$ ) that allows survival rate to be different for males than females but constrains survival rate to be the same in all years of the study.

Given the constraint of nesting and because the information-theoretic overcomes this problem and has other advantages that we'll discuss soon, we won't use the likelihood-ratio approach. Instead, we'll use the information-theoretic approach in this course for comparing models.

## Information-theoretic Approaches

In most applied ecological field studies, we don't collect data using designed experiments with true control and treatment groups and proper randomization and replication. There are philosophical problems with treating model selection as a hypothesis testing problem in such situations, i.e., in observational studies. This topic has received much attention in recent years, and we will not dwell on the issue. However, there are good arguments for treating model selection as an estimation problem. Specifically, we will focus our attention on identifying a good approximating model and in considering multiple models. The information-theoretic approach also allows us to compare models regardless of whether or not they are nested.

The information-theoretic approach addresses the tradeoff between model fit (fit favors more complex models, i.e., those with more parameters) and estimator variance (precision favors simpler models, i.e., those with fewer parameters). We will use a statistic known as Akaike's information criterion (AIC).

$$AIC = -2\ln(L) + 2k$$

The idea is to select the model with the minimum AIC. The equation has deep theoretical roots. We will see that there are modified versions of *AIC* that we can and will use. But for now the formula above will suffice. The key idea for now is that *AIC* emphasizes parsimony and makes a bias-variance tradeoff when evaluating competing models.

It is crucial when using *AIC* that all models being evaluated are fit to the same set of sample data. This may seem obvious but ... Imagine a situation where you are comparing models  $S(yr,sex)$ ,  $S(yr)$ , and  $S(sex)$  for 50 animals. Now imagine that during the trapping and radio-collaring operation that 5 animals were released before you could record the sex of the animal. In this case, the  $S(yr)$  model will be fit using 50 observations whereas the other 2 models will be fit to only 45 animals, i.e., the models are NOT all fit to the same set of sample data.

In some cases (and certainly not unusual cases!), no single model may be the clear selection, i.e., there are several models with *AIC* values that are similar and low. In such cases, model weights can be calculated based on the *AIC* scores. Such weights can be roughly interpreted as the probability that a given model is the best approximation to truth *given the dataset and the model list under consideration*. It is also possible to use the weights to obtain weighted averages of parameter estimates from across all models. The weights have appeal because they take into account the model-selection uncertainty inherent in the process.

Finally, note that in situations in which an experimental or quasi-experimental design provides a context for testing predictions based on theory, models, or both, it may be preferable to conduct a formal hypothesis test. Fortunately, in such contexts, MLE allows us to do such tests readily if they are desired. We will be delving into the details of working with multi-model inference all semester, so know that much more information is coming your way!

Simple Example of a Simple AIC Table:

| Model       | AIC   | $\Delta AIC$ | Model-Weight |
|-------------|-------|--------------|--------------|
| $S(sex)$    | 136.7 | 0.00         | 0.87         |
| $S(year)$   | 141.5 | 4.80         | 0.08         |
| $S(yr,sex)$ | 148.5 | 11.84        | 0.00         |