WILD 7250 - Analysis of Wildlife Populations
www.auburn.edu/~grandjb/wildpop

Lecture 02 – Intro to Maximum Likelihood Estimators and Information Theoretic Methods

Readings:

1. Information Theoretic Methods: model selection

   Anderson, D. R., K. P. Burnham, and W. L. Thompson. 2000. Null hypothesis testing: problems, prevalence, and an alternative. Journal Wildlife Management 64:912-923.

   Anderson, D. R., K. P. Burnham. 2002. Avoiding pitfalls when using information-theoretic methods. Journal Wildlife Management 66:912-918.

   Johnson, D. H. 1999. The insignificance of statistical significance testing. Journal of Wildlife Management 63:763-772.

Other resources:

1. Maximum Likelihood Estimators

   Azzalini, A. 1996. Statistical inference based on the likelihood. Monographs on statistics and applied probability No. 68. Chapman & Hall. New York 341 p.

   Royall, R. M. 1997. Statistical Evidence: A Likelihood Paradigm. Monographs on Statistics and Applied Probability No. 71. Chapman & Hall. New York 341 p.

2. Information Theoretic Methods: model selection

   Burnham, K. P., and D. R. Anderson. 2002. Model selection and multimodel inference: a practical information theoretic approach. 2nd ed. Springer-Verlag, New York, NY.

Maximum Likelihood Estimators

1. Binomial Sampling

Binomial sampling is characterized by two mutually exclusive events. For example the heads or tails, on or off, heads or tails, male or female, or in our case survived or died. These events are often referred to as **Bernoulli trials**. These trials have with them an associated parameter p, usually referred to as the probability of success. Thus, the probability of failure is 1-p, often referred to as q, such that $p + q = 1$.

These probabilities represent a model, in as much as they are approximations of the truth. In the case of the binomial sample, the models may be very good. When we look at more complex models used for capture-mark recapture analysis, the probabilities and the models themselves will not be so clear cut, and may not be close approximations of reality.

In this case, p is a continuous variable between 0 and 1 ($0 \leq p \leq 1$) and the estimator of p is:

$$\hat{p} = y/n$$

where y is the number of successful outcomes and n is the number of trials. This estimator is unbiased. Obviously

$$E(y) = np\,.$$

Furthermore,

$$\text{var}(y) = npq = np(1-p)$$
$$\text{var}(\hat{p}) = {pq}/{n}$$
$$\hat{\text{var}}(\hat{p}) = {\hat{p}\hat{q}}/{n} \quad .$$
$$\hat{\text{se}}(\hat{p}) = \sqrt{{\hat{p}\hat{q}}/{n}}$$

2. Important points about Binomial random variables:
   a. The n trials must be identical

      (i.e., the population is well defined e.g., 20 coin flips, 50 Kirtland's warbler nests, 75 radio-marked female Canada geese on the Copper River Delta).

   b. Each trial results in one of two mutually exclusive outcomes.

      Event or nonevent, survived or died, successful or failed, etc.

   c. The probability of success on each trial remains constant.

      Mortality rates are constant among subjects.

   d. Trials are independent events.

      Survival times are independent.

   e. y, the number of successes; is the *random variable* after n trials.

3. Binomial Probability Function and it's likelihood

$$f(y \mid n, p) = \binom{n}{y} p^y (1-p)^{n-y}$$

is read as the probability of observing y successes given n trials with the underlying probability p is ...

For example, to describe the traditional example with 10 coin flips of a fair coin (p = 0.5), 7 of which turn up heads, we would write

$$f(y \mid 10, 0.5) = \binom{10}{7} 0.5^7 (1-0.5)^{10-7}$$

evaluated numerically we arrive at

$$\begin{aligned} f(y \mid 10, 0.5) &= \frac{n!}{y!(n-y)!} \times p^y \times (1-p)^{n-y} \\ &= 120 \times 0.5^7 \times (1-.5)^{10-7} \\ &= 0.1172 \end{aligned}$$

However in reality, we usually have data (n and y), but we rarely know the parameters (p), which leads us to the likelihood function. The likelihood function is written

$$\_(p \mid n, y) = \binom{n}{y} p^y (1-p)^{n-y}$$

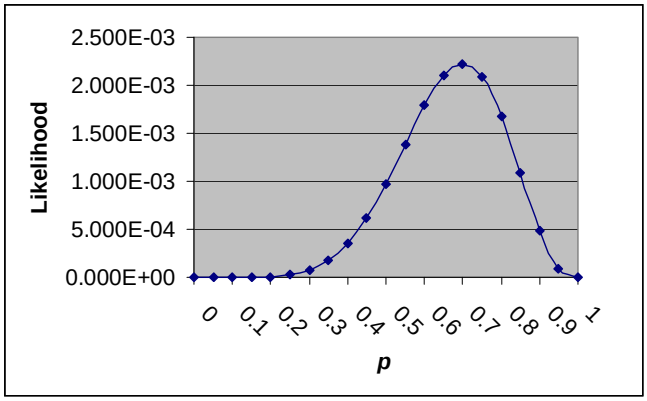
and is read the likelihood of p given n and y. The likelihood function is not a probability function. It is a positive function and $0 \leq p \leq 1$. The probability function returns the likelihood of the data given the sample size and the model (which includes the parameter estimates). The likelihood function gives the relative likelihoods for different values of the parameter given the sample size and the data.

Back to our example... We found that:

$$\binom{n}{y} = \binom{10}{7} = \frac{n!}{y!(n-y)!} = 120$$

This value is constant for our given data set and can be ignored. Thus. if we evaluate the likelihood function ignoring the above constant for various values of p we find that given our data 0.7 is the most likely value for p.

p	Likelihood
0.00	0.00000000
0.05	0.00000000
0.10	0.00000007
0.15	0.00000105
0.20	0.00000655
0.25	0.00002575
0.30	0.00007501
0.35	0.00017669
0.40	0.00035389
0.45	0.00062169
0.50	0.00097656
0.55	0.00138732
0.60	0.00179159
0.65	0.00210183
0.70	0.00222357
0.75	0.00208569
0.80	0.00167772
0.85	0.00108195
0.90	0.00047830
0.95	0.00008729
1.00	0.00000000



An alternate interpretation of likelihood is the product of the probability of each observation, and in the binomial example that probability is:

$$P(y_i) = p^f (1-p)^{1-f},$$

where f is the outcome of the event (1 or 0). Therefore:

$$\_(p \mid n, y) = \prod_{i=1}^{n} p^{f} (1-p)^{1-f}$$

Although the Likelihood function is useful, the log-likelihood has some desirable properties in that the terms are additive and the binomial coefficient does not include p. From our original formulation:

$$\ln(\_) = \ln(\_(p \mid n, y)) = \ln \binom{n}{y} + y \cdot \ln(p) + (1-p) \cdot \ln(n-y).$$

Using the alternative:

$$\ln(\_) = \ln(\_(p \mid n, y) = \sum_{i=1}^{n} \ln(p^{f} (1-p)^{1-f})$$

Once again the estimate of p that maximizes the value of ln(_) is the MLE.

4. Properties of MLEs:

as $n \Leftrightarrow \infty$

a. Asymptotically normally distributed,
b. Asymptotically minimize variance, and
c. Asymptotically unbiased.
d. Also, One-to-one transformations of MLEs are also MLEs. For example mean lifespan,

$$\hat{\overline{L}} = 1/\ln(\hat{S}),$$

is also an MLE.

Information Theoretic Methods: model selection

What's a model?

Model — An approximation of reality. A *statistical model* is a mathematical expression that help us predict a *response variable* as a function of *explanatory variable* based on a set of assumptions that allow the model not to fit exactly. These assumptions are made about the random terms in the model called *error*.

AIC—Akaike's Information Criterion

Parsimony—economy in the use of means to an end. In the context of our analyses, we strive to be economical in the use of parameters to explain the variation in data for a number of reasons.

Models of any type only approximate reality; thus we seek a good model supported by the empirical data. Thus, model selection is critical to many types of analysis. The Kullback-Leibler (Kullback and Liebler 1951) "distance," or "information" seeks to describe the difference between models and forms theoretical basis for data-based model selection. Akaike's Information Criterion or AIC (Akaike 1973) is a simple relationship between

expected Kullback-Leibler information and Fisher's maximized log-likelihood function (deLeeuw 1992). This relationship leads to a simple, effective, and very general methodology for selecting a parsimonious model for the analysis of empirical data.

Akaike (1973) demonstrated that the maximum log-likelihood is biased upward. He also demonstrated that this bias is approximately equal to K, the number of estimable parameters in the model. Thus, an approximately unbiased estimator of the relative, expected K-L information is

$$\ln(\mathcal{L})\text{-}K$$

Akaike (1973) then defined "*an information criterion*" (AIC) as

$$-2\ln(\mathcal{L})+2K$$

Thus, the model that yields the smallest value of AIC is estimated to be "closest" to reality (unknown but approximated by the model), from among the candidate models considered. If none of the models are good, AIC attempts to select the best approximating model of those in the candidate set. Thus, it is extremely important to assure that the set of candidate models is well-substantiated.

Note that AIC is only valid when comparing models fit to the same data.

1. Small sample adjustment

$$\text{AIC}_c = \text{AIC} + \frac{2K(K+1)}{n-K-1}$$

2. Quasi-likelihood (QAIC)
   a. Overdisperion

      When a lack of independence among individuals leads to a sampling variance greater than the theoretical (model based) variance, the situation is called "overdispersion." Examples include animals that mate for life and the pair behaves almost as an individual rather than as two independent "trials" or young of some species that continue to live with the parents for a period of time. Other examples include species that travel in flocks or schools. A different type of overdispersion stems results from individuals having unique parameters rather than the same parameter (such as survival probability). Overdispersion can be detected by examining model "fit."

   b. Goodness of fit

      Overdispersion can be detected by examining model "fit" much the same way you examined expected frequencies of genotypes and phenotypes in general biology and genetics labs. If overdispersion is detected then it is appropriate to apply a variance inflation factor (c) to compensate for bias. If the model fits the data perfectly then c = 1.

An adjustment to AIC that incorporates $\hat{c}$ is the Quasi-likelihood (Lebreton et al. 1992) and is calculated as:

$$\text{QAIC} = -2[ln(\mathcal{L})/\hat{c}\,]+2K$$

and for small samples:

$$\text{QAIC}_c = -2[ln(\mathcal{L})/\hat{c}] + 2K + \frac{2K(K+1)}{n-K-1}$$

We will see examples of how this procedure may influence the results of model selection and AIC weight considerably. **Note that in QAIC$_c$ *K* is increased by one because $\hat{c}$ is also estimated.**

Ranking Models: AIC differences

Because AIC values are relative the differences between AIC values,

$$\Delta\text{AIC}_i = \text{AIC}_i - \min(\text{AIC}_i),$$

are often reported rather than the values themselves. A larger ΔAICi reflects a greater distance between models and less likelihood that a model is the "best model." In general, Burnham and Anderson (1998) recommend that models with $\Delta\text{AIC} < 2$ receive consideration, models with $4 \leq \Delta\text{AIC} \leq 7$ have less support, and models with $\Delta\text{AIC} > 10$ have no support and should not be considered.

Strength of Evidence for Alternative Models

The likelihood of model i, given the data is

$$\mathcal{L}(M_i \mid x) = \exp\left(-\frac{1}{2}\Delta_i\right)$$

These can be normalized (so they sum to 1) and interpreted as probabilities:

$$w_i = \frac{\exp\left(-\frac{1}{2}\Delta_i\right)}{\sum_{r=1}^{R}\exp\left(-\frac{1}{2}\Delta_i\right)}$$

Perhaps a more useful application is in assessing the relative "strength of evidence" for a pair of competing models. For example, the relative strength of evidence for model *i* versus model *j* is estimated by:

$$w_i/w_j.$$

Model Selection Uncertainty and Multi-model Inference

Just as the w_i reflect the relative strength of evidence for model(s), the implication is that there is some degree of uncertainty in the model selection process. Indeed, the suite of models under consideration may not include the "true" model. Furthermore, several models may have $\Delta\text{AIC}_i < 2.0$. Thus, each deserved consideration when estimating the parameter(s) of interest and their precision. Unconditional parameter estimates may be calculated as the weighted average of the estimates from all candidate models.

$$\bar{\hat{\theta}} = \sum_{i=1}^{R} w_i I_i \hat{\theta}_i,$$

where *Ii* = 1 if the parameter appears in the model.

Estimates of precision based on a single model are **conditional on the selected** model and tend to overestimate precision. The w_i can be used similarly to estimate the unconditional variance of a parameter θ over a suite of models, *R*, using a sort of weighted average in the formula:

$$\hat{\text{var}}(\hat{\theta}_i) = \left[\sum_{i=1}^{R} w_i \sqrt{\hat{\text{var}}(\hat{\theta}_i \mid M_i) + (\hat{\theta}_i - \bar{\hat{\theta}})^2} \right]^2 .$$

NOTE: This concept has been extended beyond the scope of ML estimation to other types of models. For example in their book Burnham and Anderson (2002) point out that for least-squares estimation models (e.g., regression and AOV)

$$\ln(\mathcal{L}(\hat{\theta})) = \frac{-n}{2} \ln(\hat{\sigma}^2),$$

where $\hat{\sigma}^2 = \text{RSS}/n$ and RSS is the residual sum of squares $\left(\sum (\hat{\varepsilon})^2\right)$.

Hypothesis testing

It is usually preferable to use AIC in model selection procedures. When is it is appropriate to use formally examine planned comparisons among nested models Likelihood Ratio Tests (LRTs) are constructed. As the name implies:

$$\text{LRT} = -2 \ln\left(\frac{\mathcal{L}_s(\hat{\theta})}{\mathcal{L}_g(\hat{\theta})} \right)$$

where the simpler model (*s*) has fewer parameters than the general model (*g*). Recall that

$$\ln\left(\frac{A}{B}\right) = \ln(A) - \ln(B)$$

and the deviance is

$$-2\ln(\mathcal{L}_i) + 2\ln(\mathcal{L}_{sat}) .$$

Since the $-2\ln(\mathcal{L}_{sat})$ is constant for models derived from the same data

$$\begin{aligned} \text{LRT} &= -2\left(\ln\left(\mathcal{L}_s(\hat{\theta})\right) - \ln\left(\mathcal{L}_g(\hat{\theta})\right)\right) \\ &= -2\ln\left(\mathcal{L}_s(\hat{\theta})\right) + 2\ln\left(\mathcal{L}_g(\hat{\theta})\right) \\ &= \text{deviance}_s - \text{deviance}_g \end{aligned}$$

LRT is distributed approximately as χ^2 with K_s - K_g degrees of freedom (where K is the number of estimated parameters. An LRT with $P \leq \alpha$ indicates that the additional parameters in the more general modal are warranted.