Model Selection with Information Theory

"Far better an approximate answer to the *right* question, which is often vague, than an *exact* answer to the wrong question, which can always be made precise."

– John W. Tukey, (1962), "The future of data analysis." *Annals of Mathematical Statistics* **33,** 1-67.

1 Problems with Statistical Hypothesis Testing

1.1 Indirect approach:

– effort to reject null hypothesis (H_0) believed to be false *a priori*
(*statistical* hypotheses are not the same as *scientific* hypotheses)

1.2 Cannot accommodate multiple hypotheses (e.g., Chamberlin 1890)

1.3 Significance level (α) is arbitrary

– will obtain "significant" result if n large enough

1.4 Tendency to focus on P-values rather than magnitude of effects

2 Practical Alternative: Direct Evaluation of Multiple Hypotheses

2.1 General Approach:

2.1.1 Develop multiple hypotheses to answer research question.

2.1.2 Translate each hypothesis into a model.

2.1.3 Fit each model to the data (using least squares, maximum likelihood, etc.).
(fitting model ≅ estimating parameters)

2.1.4 Evaluate each model using information criterion (e.g., AIC).

2.1.5 Select model that performs best, and determine its likelihood.

2.2 Model Selection Criterion

2.2.1 Akaike Information Criterion (AIC): relates information theory to maximum likelihood

$$\text{AIC} = -2\log_e[\mathcal{L}(\hat{\theta}\,|\,data)] + 2K$$

$\hat{\theta}$ = estimated model parameters

$\log_e[\mathcal{L}(\hat{\theta}\,|\,data)]$ = log-likelihood, maximized over all θ

K = number of parameters in model

Select model that minimizes AIC.

2.2.2 Modification for complex models (K large relative to sample size, n):

$$\text{AIC}_\text{c} = -2\log_e[\mathcal{L}(\hat{\theta}\,|\,data)] + 2K + \frac{2K(K+1)}{n-K-1}$$

use AIC_c when $n/K < 40$.

2.2.3 Application to least squares model fitting:

$$\text{AIC} = n \cdot \log_e(\hat{\sigma}^2) + 2K$$

$\hat{\sigma}^2 = RSS/n$

RSS = residual sum of squares

Modification for relatively small sample size, $n/K < 40$

$$\text{AIC}_\text{c} = n \cdot \log_e(\hat{\sigma}^2) + 2K + \frac{2K(K+1)}{n-K-1}$$

2.3 Ranking Alternative Models

2.3.1 Re-scale AIC values to give best model value of 0:

$\Delta_\text{i} = \text{AIC}_\text{i} - \min \text{AIC}$

2.3.2 Use Δ_i to measure relative plausibility of each model (larger Δ_i means less plausible)

2.4 Model likelihood, given the data $\mathcal{L}(g_i \mid data)$

2.4.1 Transform Δ_i values to likelihood: $\exp(-_{1/2}\Delta_i)$

2.4.2 Normalize transformed values → "Akaike weights"

= probability that model *i* is best among alternatives considered

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

2.4.3 Relative likelihood of model *i* vs. model *j* = w_i / w_j

3 Example: Distribution of Maple seed dispersal distances

3.1 Alternative distributions:

3.1.1 Uniform Distribution:

pdf: $f(x) = \frac{1}{B-A}$ for $A \le x \le B$ shape: horizontal line

3.1.2 Binomial Distribution: 2 mutually exclusive outcomes per event

pdf: $P\{x=k\} = \binom{n}{k} p^k (1-p)^{n-k}$, for *k* = 0, 1, 2, …, n where $\binom{n}{k} = \frac{n!}{k!(n-k)!}$

mean: np var: $np(1-np)$

3.1.3 Poisson Distribution: e.g., number of trials until *k* events occur

pdf: $P(x=k) = \frac{e^{-\lambda n}(\lambda n)^k}{k!}$, for *x* = 0, 1, 2, …

mean: λn var: λn

3.1.4 Normal Distribution:

pdf: $f(x) = \frac{e^{-(x-\mu)^2/(2\sigma^2)}}{\sigma\sqrt{2\pi}}$ mean: μ var: σ^2

3.2 Fit each distribution to the data:

3.2.1 Estimate mean, variance.

3.2.2 Calculate expected frequencies, using estimated mean, var.

3.2.3 Calculate residual sum of squares (betw/ data and expected frequencies).

3.3 Calculate AIC score for each distribution.

3.4 Select model with minimum AIC

3.5 Calculate Akaike weight for selected model (and others).

Further Reading:

Anderson DR, Burnham KP. 2002. Avoiding pitfalls when using information-theoretic methods. *J. Wildl. Manage.* 66:912-918.

Anderson DR, Burnham KP, Thompson WL. 2000. Null hypothesis testing: Problems, prevalence, and an alternative. *J. Wildl. Manage.* 64:912-923.

Burnham KP, Anderson DR. 2002. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2nd ed. Springer-Verlag, New York.

Chamberlin T. 1890. The method of multiple working hypotheses. *Science*. (reprinted in 1965. *Science* 148:754-759.

Ellison AM. 1996. An introduction to Bayesian inference for ecological research and environmental decision-making. *Ecol. Appl.* 6:1036-1046.

Hilborn R, Mangel M. 1997. *The Ecological Detective: Confronting Models with Data*. Princeton Univ. Press, Princeton, NJ.