Natural Resources Data Analysis – Lecture Notes
Brian R. Mitchell

I. Week 1:

A. Welcome and Introductions

B. Review Syllabus

1. Review some specifics:

2. Written assignments are aimed towards developing a professional manuscript (thesis chapter or journal publication) describing your results.

3. Check the web site weekly for changes (and new readings if any are listed as "TBA"). I will post any changes by the end of the day on Monday for the following week's class.

4. Let me know if there are topics you wanted to see that aren't on the schedule; I will try to add topics in if there is interest.

5. First few weeks will focus on developing model sets, then quickly proceed into nuts and bolts of data analysis.

C. Data sets

1. Each participant should give a brief summary of their research question and data set, and describe the statistical approach they want to use and why.

D. Approaches to data analysis

1. I think it is worth taking some time to discuss the various approaches to testing statistical hypotheses. Let's start at a really basic level and work up.

2. What is a ***hypothesis*** (in the most general sense)?

3. A hypothesis is a tentative explanation for an observation, phenomenon, or problem.

4. What is a scientific hypothesis?

5. A scientific hypothesis is a word model that tries to explain or make a prediction based on our current understanding of a problem (a ***scientific model***).

6. What is a statistical hypothesis?

7. It is a statement about the attributes of a ***statistical model*** whose validity can be assessed by seeing how well the model matches data. Note that a statistical model is explicit, quantitative, and includes a description of uncertainty (error).

8. There should be a ***one-to-one correspondence*** between a scientific model and a statistical model. If there is a one-to-one correspondence, then learning about the adequacy of a statistical hypothesis can teach us about the adequacy of a scientific model.

9. Example: scientific model: in a sexually reproducing population with random mating and equal parental investment, natural selection should favor equal proportions of male and female

offspring. Scientific hypothesis: there are equal proportions of male and female offspring. Statistical model: the number of males (or females) in a litter of a given size is binomially distributed. Statistical hypothesis: θ, the probability of a male offspring, = 0.5.

10. The next step is to ***collect data*** using an appropriately designed and thought-out strategy.

11. Then we need to ***evaluate*** whether the data are consistent with our statistical hypothesis.

12. What are the ***potential approaches*** we could use to evaluate a statistical hypothesis?

13. The general approaches are ***Frequentist***, ***Bayesian***, and ***Likelihood***. Let's explore each in turn.

14. Frequentist

a) There are a couple of related approaches to classical hypothesis testing; they are all considered "***frequentist***" approaches. Why? (Hint: has a lot to do with the convoluted way you were taught to talk about p-values in multivariate stats)

b) These approaches always consider the frequency with which your data or more extreme data would be collected, IF you conducted many replicate experiments.

c) Fisherian hypothesis testing (1956)

(1) Construct a statistical null hypothesis (e.g. $\theta = 0.5$).
(2) Choose an appropriate distribution (e.g. binomial distribution) or test statistic (e.g. t statistic).
(3) Collect the data with random samples.
(4) Determine the p value (probability of obtaining the value or one more extreme), assuming the null hypothesis is true.
(5) Reject the null hypothesis if p is small (and always report the p value as a "strength of evidence" measure).
(6) Fisher recommended a significance level of 0.05, but later argued that the significance level should depend on the circumstances.

d) Neyman-Pearson hypothesis testing (1933)

(1) Similar to Fisher's approach, but:
(2) Set significance level in advance, and interpret it as the proportion of times the null would be improperly rejected given many replicates and a true null hypothesis.
(3) Explicitly incorporate an alternative hypothesis; this must be true if the null is false.
(4) p value is not a strength of evidence; its only use is in deciding to accept or reject the null hypothesis.
(5) Focus on Type I and Type II errors, as well as power of tests.

e) A hybrid approach is common today

(1) This is essentially the Neyman-Pearson approach, but with the view that p values are a strength of evidence (e.g. significant, very significant (0.01), and highly significant (0.001).

f) Critique of the frequentist approach

(1) What are the problems with the frequentist approach?

(2) ***Dependence on sample size***: larger sample sizes are more likely to produce a significant result. This is not such a big deal if you use power analysis to set the sample size a priori.
(3) ***Unobserved data***: p-values represent the probability of the data observed as well as more extreme data (i.e. data not observed). So when what we want is the probability of the data, what we get is the probability of the data or data that is more extreme.
(4) ***Probability of the data***: What we want to know is $P(H_0|data)$, but what we get is $P(data|H_0)$ or, more accurately, we get P(data or data more extreme$|H_0$).
(5) ***Trivial null*** hypothesis: the null hypothesis is almost always trivial and logically false. BUT, rejecting the null is not important because it was believed; it is important because it indicates the presence of an effect worth reporting and investigating. Also, nulls do not have to be stated as "no effect".
(6) ***arbitrary significance*** levels: alpha level is arbitrary. (Even when it is chosen before-hand, and is not necessarily 0.05, it is still arbitrary).
(7) No way to include ***prior knowledge*** into the analysis.
(8) ***Suggestions*** for using the classical approach: Focus on ***effect sizes*** and confidence intervals. If you can't calculate your sample size in advance (with an a priori power analysis), then take care to separate statistical significance from ***biological significance***.

15. Bayesian

a) The Bayesian approach avoids many of the problems with the frequentist approach, but the main lure of Bayesian analysis is that it 1) provides a way of dealing with the frequentist issue related to ***probability of the data***. In other words, classical analysis tells us $P(data|H_0)$, but we are interested in learning $P(H_0|data)$; and 2) Bayesian analysis also allows the incorporation of ***prior knowledge***.

b) Bayesians believe that analyses and decisions should be made on the basis of the observed data, not on the data that might have been observed in a series of hypothetical experiments. The result of a Bayesian analysis is a changed degree of belief in the hypotheses being investigated.

c) Bayes Theorem

(1) Who knows Bayes theorem?
(2) ***Bayes theorem***: $P(H_1|data) = P(data|H_1)P(H_1)/P(data)$
(3) OR: ***posterior probability*** of H_1 = (***likelihood*** of observing the data given H_1)*(unconditional ***prior probability*** of H_1, taking into account existing knowledge)/(mean of the likelihood function, which serves to ***standardize*** the area under the posterior probability curve so it equals 1).
(4) OR: the posterior probability is proportional to the likelihood times the prior probability.
(5) Prior probability distributions can take two forms: 1) prior ignorance (a ***non-informative*** distribution). This helps overcome the potential subjectivity in a bayesian analysis. Most common is a uniform distribution. 2) substantial prior knowledge, represented by an ***informative prior*** probability distribution (e.g. a normal or beta distribution).

d) An example of Bayesian parameter estimation

(1) Estimating sulfates (μmol/L) in streams in NY
(2) Assume a normal distribution for the prior and the data.

(3) Prior: mean = PM = 50, variance = PV = 44
(4) Sample: mean = SM = 61.92, variance = SV = 37.46, n = 39
(5) Posterior variance = 1/(1/PV + n/SV) = 0.94
(6) Posterior mean = posterior variance*(PM/PV + SM*n/SV) = 61.67
(7) The posterior mean and probability is a weighted average of the prior and the data.
(8) What happens ***if PV is more certain*** (smaller variance)?
(9) If the prior is more certain (less variance), it has more of an effect on the posterior probability. For example, PV was 10, posterior mean = 60.88 and posterior variance = 0.88.
(10) What if the PM is lower?
(11) A prior with a more distant mean will also have a greater effect.
(12) What if the prior is uninformative?
(13) If the prior is uniform (uninformative), then the posterior mean and variance will simply equal the sample mean and variance.

e) Bayesian hypothesis testing

(1) There is no formal accept/reject decision framework; simply attach greater or lesser favor to the alternatives based on the shape of the posterior distributions.
(2) Can formalize hypothesis testing by calculating P(H|data) for each hypothesis, then calculating a ***posterior odds ratio***: $P(H_A|data)/P(H_B|data)$
(3) The posterior odds ratio also = "***Bayes factor***" * the prior odds ratio. If the two hypotheses were considered equally likely a priori, then the Bayes factor = the posterior odds ratio
(4) The Bayes factor and the posterior odds ratio are measures of the ***weight of evidence*** in favor of H_A and against H_B. The magnitude of the Bayes factor is used as evidence in favor of a hypothesis.
(5) A simpler alternative to the Bayes factor is the ***Schwarz Criterion*** (or Bayes Information Criterion, or ***BIC***); this approximates the log of the Bayes factor and is easy to calculate.

f) Problems with Bayesian analysis

(1) What are the problems with Bayesian analysis?
(2) Using prior information amounts to "personal opinion" and is ***inherently biased***.
(3) Determining the appropriate ***statistical form for prior information*** is complicated.
(4) Practical application of a Bayesian analysis is ***complicated***. Although computer programs are getting better (e.g. WinBUGS for MCMC analyses), complex models simply cannot be evaluated or estimated with a Bayesian approach due to the difficult calculus involved.

16. Likelihood

a) Background

(1) Likelihood inference takes the evidence that the observed data provide about the hypothesis and represents it as a likelihood function (likelihood of the data, given the hypothesis).
(2) ***Likelihood inference*** is about relative measures of evidence of support between competing hypotheses, and the focus is on the likelihood ratio (which is the relative strength of evidence provided by the data supporting H_1 compared to H_2.

(3) The ***likelihood function*** of a parameter or variable (e.g. the proportion of heads in a coin toss) can be thought of as a graph of the relative chance of observing a given value (on the y axis) against all possible values of the parameter (on the x axis).
(4) The ***Law of Likelihood*** say that if hypothesis A implies that the probability of a random variable X taking value $x = \rho_a(x)$, while hypothesis B implies that the probability of a random variable X taking value $x = \rho_b(x)$, then the observation that X = x is evidence supporting A over B if $\rho_a(x) > \rho_b(x)$, and the ***likelihood ratio*** $\rho_a(x)/\rho_b(x)$ measures the strength of that evidence.
(5) As a ***rule of thumb***, likelihood ratios below 8 are considered weak evidence, between 8 and 32 is moderate evidence, and above 32 is strong evidence (Royall 2004).

b) Example

(1) We are given a coin that we suspect is biased towards excess heads.
(2) We toss the coin n = 20 times, and get x = 12 heads.
(3) Hypothesis A is that the coin is unbiased ($\pi = 0.5$), and Hypothesis B is that heads will occur 60% ($\pi = 0.6$) of the time. We choose this value for hypothesis B knowing it will yield the maximum likelihood ratio based on the data collected.
(4) We use the binomial distribution: $P(x) = \frac{n!}{x!(n-x)!}\pi^x(1-\pi)^{n-x}$
(5) $\rho_a(x) = 0.1201$, and $\rho_b(x) = 0.1797$, so the likelihood ratio is 1.50
(6) ***How strong*** is the evidence that this coin is biased?
(7) What if n = 150?
(8) $\rho_a(x) = 0.00324$, and $\rho_b(x) = 0.06637$, so the likelihood ratio is 20.50
(9) ***How strong*** is the evidence that this coin is biased now?

c) Misleading evidence

(1) Note that regardless of how strong the evidence is, it can still be ***misleading***. In other words, there is still a chance that the coin in the example was not really biased, even though there is strong evidence that it is. The interpretation (that the coin was biased) is still correct; it was the evidence that was misleading.
(2) The ***maximum probability of misleading evidence*** cannot exceed 0.021 when the likelihood ratio exceeds 8, and cannot exceed 0.004 when the likelihood ratio exceeds 32 (Royall 2004, citing Royall 1997).

d) Likelihood-based hypothesis testing

(1) ***Experimental design*** should consider the minimum sample size at which the probability of generating weak evidence for distinguishing between the hypotheses is low. Note that if the probability of generating weak evidence is low, the probability of misleading evidence will be even lower.
(2) ***Observed data*** is assumed to fit some underlying probability model (as in frequentist methods)
(3) The ***likelihood ratio*** provides an explicit and objective measure of the strength of the statistical evidence.
(4) There is no dependence on a particular ***stopping rule***; there is no reason not to collect additional data if the likelihood ratio indicates weak data; researchers are encouraged to examine the likelihood functions of their data and adjust the sample size accordingly. This is of course absolutely forbidden in the frequentist approach.

e) Problems with likelihood analysis

(1) ***Arbitrary*** levels of importance.
(2) Does not incorporate ***prior information***.
(3) Any other thoughts?

17. Relationships between approaches

a) Likelihood has some aspects of the frequentist approach, because likelihood ratios follow a χ^2 distribution (actually 2*ln(LR) is the correct statistic) and can be tested using frequentist methods

b) Likelihood also has aspects of the Bayesian approach. With a uniform prior, the Bayesian posterior probability distribution has an identical shape to the likelihood function.

18. What approach should be used, and when?

a) When should the ***classical approach*** be used?

(1) ***Strict experiments*** with control and treatment

b) When should a ***Bayesian approach*** be used?

(1) Any situation where you want to incorporate ***prior knowledge***
(2) These situations can include ***model selection*** and determination of ***effect sizes***
(3) Particularly well suited to ***learning algorithms*** (e.g. neural networks)

c) When should a ***likelihood approach*** be used?

(1) Natural experiments
(2) Observational experiments
(3) Determination of effect sizes
(4) Model selection

E. Approaches to model selection and averaging

1. In this class we are concerned with ***model selection***. The model sets can range from alternative hypotheses about ***historical events***, to hypotheses about ***processes*** that generated a data set, to hypotheses about which ***parameters*** are most important in a data set.

2. We might be interested in ***predicting*** future data, ***understanding*** the existing data, or ***estimating*** effect sizes.

3. Frequentist methods

a) How would a ***frequentist*** go about selecting the best model?

b) In general, the frequentist approach is not a good strategy for model selection, primarily because: 1) hypothesis tests between models are not independent, and 2) there are serious problems with the probability of Type I error due to multiple testing of the same data.

c) Stepwise

(1) ***Model set*** would be a nested set ranging from an intercept-only model to the most general model (i.e. including all parameters plus important interactions).

(2) ***Stepwise parameter selection*** is used to compare two models at a time. Can be forwards stepwise (begin with no parameters) or backwards stepwise (begin with full model). At each step, evaluate whether adding the most important excluded parameter increases model fit, and evaluate whether removing the least important included parameter decreases model fit. Continue until you get to the point where adding any one of the remaining excluded parameters does not improve the fit, and removing any one of the included parameters decreases the fit.
(3) Selection is generally based on the likelihood ratio ***F-test***.

d) All subsets

(1) Same ***model set*** as above
(2) Calculate some statistic (e.g. adjusted R^2, AIC, BIC) for ***every model***, and pick the model with the best value.
(3) NOTE: using AIC and BIC in this situation is ***essentially a likelihood approach*** to model selection; it is valid but considered weak (exploratory). Adjusted R^2 functions poorly as a model selection criteria.

4. Bayesian approach

a) How would a ***Bayesian*** go about selecting the best model?

b) Develop a ***model set***; would probably not include all possible models.

c) Explicitly consider the ***prior information*** about which model(s) are most likely.

d) Calculate the ***Bayes factor*** for each model (or its approximations, ***BIC*** or ***AIC***, depending on your model priors), and select the best model or use model averaging.

5. Likelihood approach

a) How would you go about model selection using a ***likelihood*** approach?

b) Develop a ***model set***; would probably not include all possible models.

c) For each model, calculate the ***likelihood*** of the data, given the model.

d) Calculate ***AIC*** or some other information criterion, and select the best model (if evidence overwhelmingly supports it) or use model averaging (if multiple models are supported by the evidence).

e) This is the ***Information-Theoretic*** approach to model selection and averaging, and is based on the concept of ***Kullback-Leibler*** distance

F. Discuss specifics from the readings

1. Let's take the rest of the class period to discuss the readings for this week, which provided background information for the information-theoretic approach. We'll start with Johnson and Omland before going into B&A, which is more technical.

2. Johnson and Omland:

a) In my view, the main points are:

b) ***Benefits of model selection***: 1) Competing models are ***compared*** to one another by evaluating the relative support for each. 2) Models can be ranked and weighted, thereby providing a ***quantitative*** measure of relative support for each competing hypothesis. 3) ***Model averaging*** can be used to make robust parameter estimates and predictions.

c) ***Steps to model selection***: 1) ***Articulate*** a reasonable set of competing hypotheses as models. 2) ***Fit*** the models to the observed data. 3) Examine the ***goodness-of-fit*** of the most heavily parameterized (i.e. global) model in the candidate set. 4) ***Select*** the best model or use model averaging.

d) ***Model averaging*** eliminates model selection bias and accounts for model selection uncertainty.

e) ***When*** should model selection be used? Model selection is well suited for making inferences from observational data, especially when data are collected from complex systems or when inferring historical scenarios where several different competing hypotheses can be put forward.

f) ***Caveats***: 1) Inferences derived from model selection ultimately ***depend on the models included*** in the candidate set. A bad candidate set is a big problem! 2) ***Models should be plausible***. If a model is to carry biological meaning, rather than mere statistical significance, then its predictions and parameter estimates must be biologically plausible. 3) ***When*** is it appropriate to use model selection, and when is it appropriate to use designed experiments and inferences based on significance tests?

3. B&A, Chapter 1:

a) In my view, the main points are:

b) 1.2.1: ***Given an appropriate model***, the maximum likelihood method objectively estimates parameters and the sampling covariance matrix. This is why model fit is important; if the model doesn't fit, the estimates are biased!

c) 1.2.2: Results obtained using ***Ordinary Least Squares*** (OLS) methods can be converted to Maximum Likelihood (ML) estimates.

d) 1.2.4: ***Candidate model sets***: Building candidate models is a subjective art. Need deep thought and early exploratory data analysis. All models must be biologically plausible. Need many more cases than variables! Aim for a model set with 4-20 models.

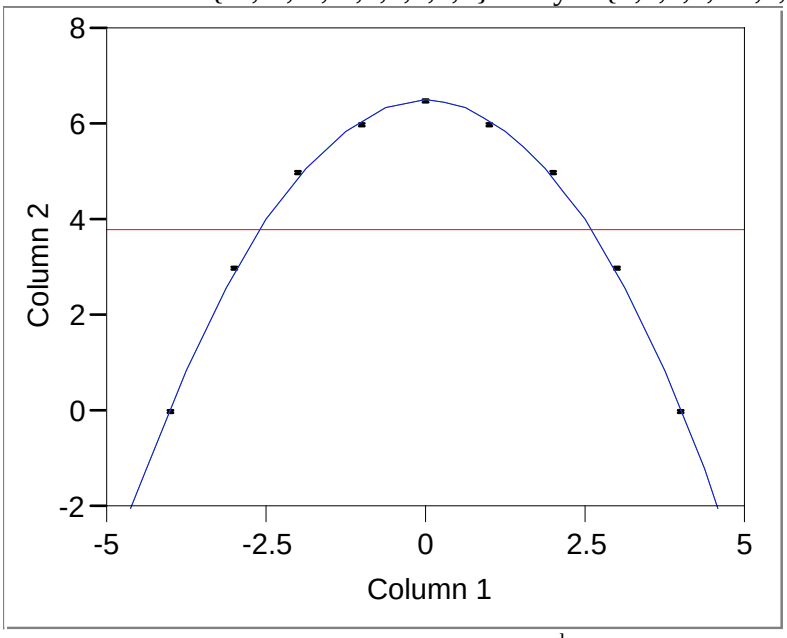
e) 1.2.5: ***No true model***; only good approximations.

f) 1.3.6: ***Global model*** should have all factors and variables thought to be important. If the global model fits data, a selected model that is more parsimonious will also fit the data.

g) 1.4.2: ***Parsimony***: use the fewest parameters that represent the data. There is a ***tradeoff*** variance and bias that results from fitting a model. ***Underfitting*** misses important structure and biases parameter estimates; you will miss important treatment effects. ***Overfitting*** produces models with very little bias, but have large estimates of sampling variance because of lack of precision in parameter estimates. These models also have spurious treatment effects.

h) Example of parsimony:

(1) Given data: x = {-4,-3,-2,-1,0,1,2,3,4} and y = {0,3,5,6,6.5,6,5,3,0}



(2)

(3)	Linear	Quadratic	3rd Order
(4) Int ± SE:	3.83 ± 0.89	6.52 ± 0.058	6.52 ± 0.063
(5) x ± SE:	0.00 ± 0.34	0.00 ± 0.015	0.00 ± 0.042
(6) x^2 ± SE:		-0.40 ± 0.006	-0.40 ± 0.007
(7) x^3 ± SE:			0.00 ± 0.003

(8) The linear model is clearly biased; the intercept is estimated at 3.83 instead of 6.5. The quadratic is the best fit; moving to the 3rd order model does not change the parameter estimates (doesn't lead to bias), but it adds an unimportant parameter and increases the SE for all parameters.

i) 1.5: ***Data dredging***: Dredging leads to overfitted models that perform much more poorly than summary statistics suggest.

j) 1.5.1: ***What constitutes dredging***? Dredging includes exploratory data analyses, also includes fitting all possible models. Dredging invalidates statistical tests and estimates of precision. Can dredge AFTER the initial a priori phase, but need to explain that this was done.

k) 1.6: ***Model selection bias***: Data-based model selection will bias estimates of model parameters. The bias is often severe.

l) 1.7: ***Model selection uncertainty***: Caused when the same data is used for model selection and parameter estimation. There can be substantial uncertainty in model selection, especially when there are tapering effects.

G. References for this class

1. Burnham, K. P. and D. R. Anderson. 2002. Model Selection and Multimodel Inference. 2nd edition. Springer, New York, New York, USA. pp. 1 – 48.

2. Johnson, J. B. and K. S. Omland. 2004. Model selection in ecology and evolution. *Trends in Ecology and Evolution* 19(2): 101-108.

3. Lewin-Koh, N., M. L. Taper, and S. R. Lele. 2004. A brief tour of statistical concepts. *In:* The Nature of Scientific Evidence. M. L. Taper and S. R. Lele, editors. The University of Chicago Press, Chicago, IL, USA. pp. 3-16.

4. Quinn, G. P. and M. J. Keough. 2002. Experimental Design and Data Analysis for Biologists. Cambridge University Press, Cambridge, UK. 537 pp. (Chapters 1 – 3)

5. Royall, R. 2004. The likelihood paradigm for statistical evidence. *In:* The Nature of Scientific Evidence. M. L. Taper and S. R. Lele, editors. The University of Chicago Press, Chicago, IL, USA. pp. 119-152.