Model Based Statistics in Biology.

Part II. Quantifying Uncertainty.

Chapter 7.4 Parameter Estimates

Quotes p2 to transparency

or find 4 students to take the parts of each source, in class

on chalk board

ReCap Part I (Chapters 1,2,3,4)
Quantitative reasoning: Example of scallops, which combined models (what is the relation of scallop density to substrate?) with statistics (how certain can we be?)

ReCap (Ch5)
Data equations summarize pattern in data as a series of parameters (means, slopes)

ReCap (Ch 6)
Frequency distributions are a key concept in statistics.
They are used to quantify uncertainty.

ReCap (Ch 7)
Inferential statistics are a logical procedure for making decisions when there is uncertainty due to variable outcomes.
Hypothesis testing is concerned with making a decision about an unknown population parameter.
Estimation is concerned with the specific value of an unknown population parameter.

Today: Parameter Estimates

Wrap-up
We use well established formulas to make the "best" estimate of a parameter.
The most common parameters in biology are
means
slopes
proportions, odds, and odds ratios

Hypothesis testing or estimation ?

Introductory courses in statistics present hypothesis testing based on the H_A / H_o logic developed in the first half of the 20th century by Fisher and Neyman. A series of quotes will serve to illustrate the history of thinking concerning experiments and hypothesis testing in the 20th century.

If your experiment needs statistics, you ought to have done a better experiment. --Ernest Rutherford (1871-1937)

Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis --R.A. Fisher 1935

Everyone will have his own pet assortment of flotsam; mine include most of the theory of significance testing, including multiple comparison tests, and non parametric statistics.
--John Nelder, 1971

Elementary statistics courses for biologists tend to lead to the use of a stereotyped set of tests with too much emphasis on hypothesis testing and too little emphasis on parameter estimation.
--M.J. Crawley 1993

Parameters.

Formal models (equations) consist of variable quantities and parameters
Parameters have a fixed value in a particular situation.
Parameters are found in
- Functional expressions of causal relations.
- Statistical or empirical functions (causal relation unknown).
- Theoretical frequency distributions (normal, poisson, etc).

We have used several kinds of models so far.
- All of these consist of variables and parameters.
  - Variables take on any value.
  - Parameters have a fixed value (in any one situation)
- These parameters are obtained from data by <u>estimation</u>

Parameters.

Parameters are used in variety of ways. They are used describing the functional relation of one quantity to another. They are used in describing pattern, such as an empirically derived relation between quantities. Parameters are used in describing frequency distributions, such as the normal or binomial distribution.

I. Functional relationships. A direct causal relation is implied.

Examples:

Scallop density as a function of substrate roughness (Lec2)
Estrogen levels in tamarin mothers (Lab 2)
Infection rate of snails by parasites (Lab 2)

2. Statistical or empirical relationships. *Y* is function of *X*, *Y* can be calculated from *X*, but *X* does not necessarily cause *Y*

Examples:

Leg lengths of water bugs *Notonecta* as function of body size (Lab 2)
Fish catch from lakes as function of the morphoedaphic index *MEI*

3. Theoretical frequency distributions.

Any theoretical distribution (probability density function pdf) will have one or more parameters.

All of these models consist of variables and parameters

1. Functional relationship. Scallops density: $M_{scal} = k_1$ if $R = 5$ or 6
   $M_{scal} = k_2$ if R not equal 5 or 6

   Variables are

   M_{scal} (kg caught per unit area of seafloor)

   R = sediment roughness from 1 (sand) to 10 (cobble)

   Parameter is k, which has two values (two means) k_1 and k_2

2. Statistical relationship. Morphoedaphic equation: $M_{fish} = 1.38\ MEI^{0.4461}$

   Variables are

   M_{fish} = kg ha^{-1} yr^{-1} (fish caught per year from lakes)

   MEI = ppm m^{-1} (dissolved organics / lake depth)

   Parameters are

   0.4461 is slope relating pM to MEI on a log-log plot

   1.38 kg ha^{-1} yr^{-1} $ppm^{-0.4461}$ $m^{0.4461}$

3. Frequency distribution: Normal distribution

   Parameters are:

   Examples: Normal distribution $pdf = f(Y) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-[(Y-\mu)^2/2\sigma^2]}$

   where the variable is $Y = \sigma X + \mu$ = outcomes with measured units,

   X is a random variable with no units

   parameters are μ (mean) and σ (standard deviation).

   Forms of each are $n \cdot P(X = x)]$ $f(x) = P(X = x)]$ = pdf

   $n \cdot P(X \leq x)]$ $F(x) = P(X \leq x)]$ = cdf

   X is a random variable. x is the counter (x-axis).

   Other examples. Binomial $f(x) = \frac{n!}{x!\,(n-x)!}\, p^x (1-p)^{n-x}$

   where x is a random variable (number of successes)

   parameters are n = number of trials, p = success per trial.

   Poisson (1 parameter, success rate p).

Parameter Estimates

1. Scallop density. The model of scallop density M_{scal} (kg/tow) was

$M_{scal} = \mu_1$ if sediment roughness $R = 5$ or 6

$M_{scal} = \mu_2$ if sediment roughness R not equal 5 or 6

We have no theoretical model from which to calculate μ_1 or μ_2
Hence the estimates of these two parameters are taken as the sample means based on n = 28 tows

M_{scal} = mean($M_{R=5,6}$) n = 13

M_{scal} = mean($M_{R \text{ ne } 5,6}$) n = 15

These sample means are calculated as shown above.

```
Data in leg1a.dat,
analysis in ctchclr3.out
```

2. Ryders's morphoedaphic equation.

Ryder's model: $pM = \alpha MEI^{\beta}$

The parameter β is a slope on a log-log plot of catch pM versus MEI
The parameter α is the intercept in this plot.

$$\ln(pM) = \ln(\alpha) + \beta \ln(MEI) \quad \text{population}$$

$$\hat{Y} = \hat{a} + \hat{\beta}_{MEI} \ln(MEI) \quad \text{sample}$$

The sample slope $\hat{\beta}_{MEI}$ is an estimate of the population parameter β_{MEI}
The sample intercept is an estimate of the population parameter $\ln(\alpha)$
The method of least squares is used to obtain the "best" estimate of these parameters.

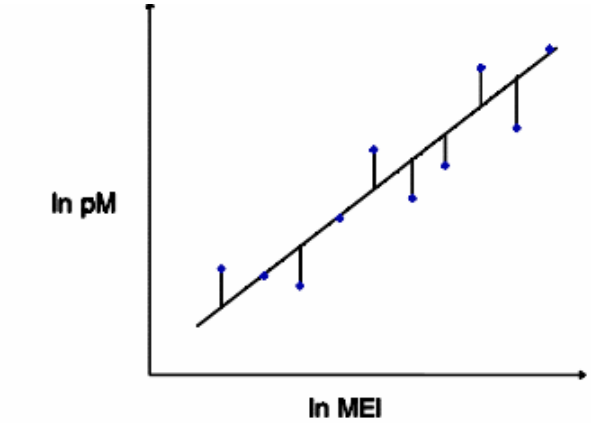

Figure L12b. Fish catch from lakes (log scale) versus Ryder's morphoedaphic index MEI (also log scale).

$$\hat{\beta}_{MEI} = Cov(Y,X) \,/\, Var(X)$$

$$Cov(Y.X) = (n-1)^{-1}\Sigma(Y-\bar{Y})(X-\bar{X})$$

$$\hat{\beta}_o = mean(Y) = \bar{Y}$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta}_{MEI} \cdot \bar{X}$$

```
Picture showing that
is the line that
minimizes the squared
vertical deviations
from the line
```

Parameter Estimates (continued)

3. The normal distribution
First, an estimate of the population mean μ

The estimate is $\bar{Y}$, which is calculated as $\bar{Y} = n^{-1}\Sigma Y$

This is an estimate of the mean value of the entire population, also called the expected value μ = E(Y)

The mean value of a sample drawn from the population is called the sample mean $\bar{Y}$

The sample mean is the "best" estimate (in the statistical sense) of the population mean μ. "best" means "minimum deviation." It is not necessarily the most accurate.

If the sample is not a representative sample of the population from which it is drawn, will this estimate be an accurate estimate of the true value μ ?

Next, an estimate of the population variance σ^2. This parameter describes the degree of spread of the normal distribution.

The estimate of the parameter σ^2 is s^2, calculated as $s^2 = n^{-1}\Sigma(Y - \bar{Y})^2$

The variance of the population is the mean squared deviation from μ

$$E(Y - \mu)^2 = \text{var}(Y) = \sigma^2$$

The variance of the sample is s^2, an estimate of σ^2. The sample variance s^2 is the "best" estimate of the variance of the population σ^2.

If the sample is not representative of the population, will this estimate s^2 be an accurate estimate of the true value σ^2 ?

$$P(Y = \mu) = (\sigma\sqrt{2\pi})^{-1}$$

Finally, an estimate of P(Y=μ).
This parameter describes the maximum frequency, which will occur at $Y = \mu$. An estimate of P(Y=μ) is

$$P(Y = \mu) = (s\sqrt{2\pi})^{-1}$$

where s is the standard deviation, an estimate of the parameter σ

The expected value of the relative frequency of Y takes on values less than P(Y=μ) if Y is greater or less than the population mean μ

Statistical Inference: Estimation

There are two categories of statistical inference:

Hypothesis testing is concerned with making a decision about an unknown population parameter.

Estimation is concerned with the specific value of an unknown population parameter.

Estimates of population parameters are made from samples.

An analytic formula is often used to make an estimate
- An example is the formula for the mean
  - (this minimizes the squared deviations of the data from the mean)
- An example is the formula for the slope of a regression line
  - (this minimizes the squared deviations of the data from the line)

These are the two most common formulae for making estimates, but they are by no means the only formulae. For example, the formula for the maximum frequency for normal data is

$$P(Y = \mu) = (\sigma\sqrt{2\pi})^{-1}$$

Estimates can also be made with an iterative procedure, rather than applying an analytic formula. Iteration continues until some criterion is reached.
The most widely accepted criterion is maximum likelihood. The criterion is that the likelihood of the parameter, given the data, be as large as possible. Commonly, this estimate is obtained iteratively by minimizing a deviance. The most common deviance is the sum of the squared deviations of the data from the model. Another common deviance is the G-statistic, for counts that arise from Poisson or binomial processes.

To evaluate our estimate, we need some measure of uncertainty. Usually this takes the form of a confidence limit for the parameter.

A Confidence limit consists of two values between which we have a specified level of on confidence (say, 95%) that the population parameter lies.

This course will demonstrate parameter estimation and confidence limits as an alternative to the machinery of hypothesis testing.