

Computer intensive methods: introduction

November 27, 2006

Outline

- ▶ What are computer intensive methods and why are they needed?
- ▶ Maximum likelihood method
- ▶ Local smoothing functions
- ▶ Bootstrap and jackknife
- ▶ Randomization (permutation) and Monte Carlo
- ▶ Bayesian methods (Petter)

What are computer intensive methods and why are they needed?

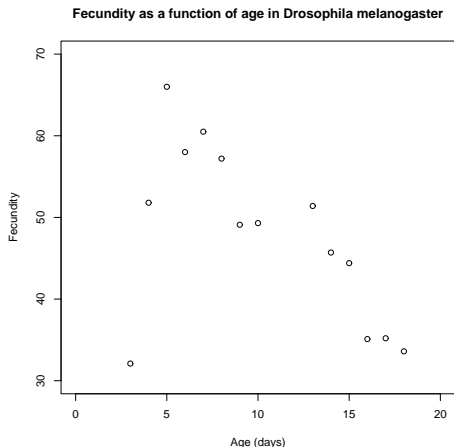
- ▶ Computer intensive methods (CIM) involve an iterative process and therefore, cannot readily be done except on a computer.
- ▶ CIM allow us to apply a statistical analysis in situations, where more “traditional” methods fail or when no analytical solutions exist
- ▶ Great efforts should be put into making the experimental design amenable to traditional methods, since these have both well-understood statistical properties and are easily carried out given the available statistical programs

Maximum likelihood method

- ▶ Basis for most of the parametric statistics taught in elementary statistics courses
- ▶ Principle of **maximum likelihood**: we select the value of a parameter that most likely (with highest probability) can be thought of having produced the data we have
- ▶ Sometimes analytical solutions can be obtained but in more complex cases the solution can only be obtained numerically
→ **CIM**

Maximum likelihood method: An example

Let us consider the relationship between age (x) and the fecundity function (f). For many animal species it looks like



Maximum likelihood method: Example continues

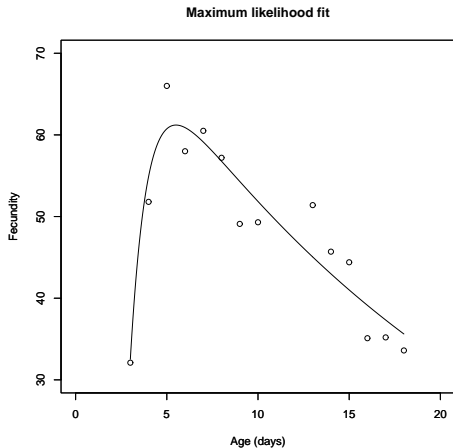
- ▶ The relationship can be represented by the model

$$f(x) = M(1 - \exp(-k(x - t_0)))\exp(-bx)$$

where M , k , t_0 and b are parameters

- ▶ Using the principal of **maximum likelihood**, it can be shown that the best estimates of the four parameters can be found by minimizing the residual sums of squares.
- ▶ This cannot be done analytically but can be done numerically
→ **CIM**

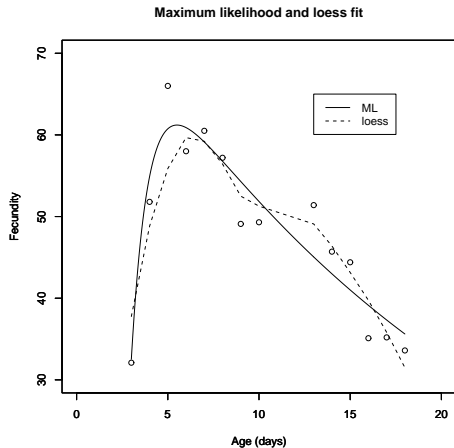
Maximum likelihood method: Example continues



Local smoothing functions: Example continues

- ▶ We may not be able to find a simple function that adequately describes the data
- ▶ We can use so-called **local smoothing** techniques (piece-wise fit through the data, keeping the fitted curve continuous and relatively smooth)
→ CIM

Local smoothing functions: Example continues



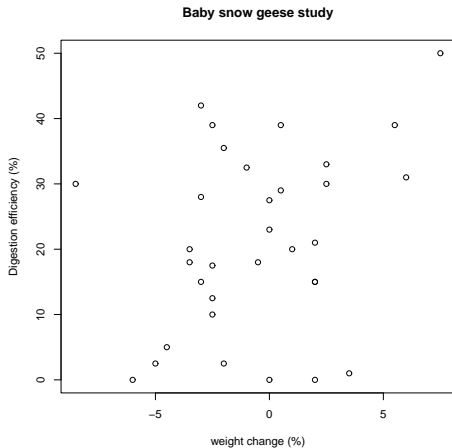
Bootstrap and jackknife

- ▶ Computer intensive methods of statistical inference (point estimation, standard error, confidence interval, hypothesis testing)
- ▶ Can be used to analyze and understand complicated data sets

Bootstrap and jackknife: An example

Problem: Feeding habits of baby snow geese were investigated. 33 goslings were without food until their guts were empty and then, they were allowed to feed for 6 hours on a diet of plants. The change in the weight of the gosling after 2.5 hours was recorded as a percentage of initial weight. Digestion efficiency (measured as a percentage) was also recorded. Correlation between weight change and digestion efficiency?

Bootstrap and jackknife: Example continues



Bootstrap and jackknife: Example continues

- ▶ The computed correlation coefficient is $\hat{r} = 0.309$. How accurate is this estimate?
- ▶ Compute the standard error (s.e.) of the correlation coefficient. If it is small, the estimate is accurate.
- ▶ There is no formula for the s.e. of the correlation coefficient and therefore, it is hard to assess the accuracy of the estimate
→ CIM: bootstrap or jackknife

Bootstrap and jackknife: Example continues

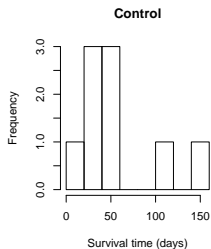
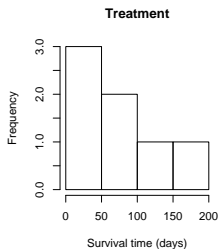
Bootstrap: We draw a **bootstrap sample**, a random sample of size n (here 33), with replacement from the observed pairs of values, and calculate the correlation coefficient from this sample, say \hat{r}_1^* . This is repeated a number of times (N). The **bootstrap** estimate of the standard error of the correlation coefficient is then

$$se(\hat{r}) = \sqrt{\sum (r_i^* - \text{mean}(r_i^*))^2 / (N - 1)}$$

Jackknife: We draw a **jackknife sample** by leaving out one pair of observations, and compute the correlation coefficient from this sample. An observation is left out at the time giving us n (here 33) samples of size $n - 1$ (32). The **jackknife** estimate of standard error of the correlation coefficient is the same as above but the bootstrap estimates \hat{r}_i^* are replaced by estimates from the **jackknife** samples.

Randomization: An example

Experiment: 7 out of 16 mice were randomly selected to receive a new medical treatment, while the remaining 9 were assigned to the non-treatment (control) group. The treatment was intended to prolong survival after a test surgery.



Randomization: Example continues

- ▶ We want to test for a difference between the mean survival time in the treatment group and that in the control group in the mice example, i.e

$$H_0 : \mu_t = \mu_c \quad \text{against} \quad H_1 : \mu_t > \mu_c$$

- ▶ This can be done by
 - ▶ The parametric t -test: tests under the assumption of normal distribution of the means; does not work for small sample sizes
 - ▶ The non-parametric Mann-Whitney test: tests differences in central tendency (usually median), not differences in the mean
 - ▶ **Randomization (permutation)** can be used in the case of both normal and non-normal distributions, subject to the restriction that a difference in means is the only difference between the two distributions
- CIM

Randomization: Test construction

- (1) Compute the difference between the two means in the data
- (2) Distribute the 16 observations at random between the treatment group (7) and the control group (9), and compute the difference between the two means
- (3) Repeat (2) a large number of times, say N
- (4) Count up the number of cases, say n , in which the absolute difference between the two means from the randomized data set is larger than that from the original data
- (5) Estimate the probability, P , under the null hypothesis of obtaining a deviation as large or larger than observed as

$$P = \frac{n + 1}{N + 1}$$

- (6) If $P < 0.05$ (for example), the observed difference is declared to be unlikely under the null hypothesis and the null hypothesis is rejected

Randomization: Conclusions

- ▶ Useful for small samples which cannot be assumed to come from a normal distribution
- ▶ Can be used for hypothesis testing and constructing confidence intervals
- ▶ Differs from bootstrap in that sampling is done without replacement
- ▶ As test statistic (here difference between the two means) any statistic can be used

Monte Carlo methods

- ▶ Both bootstrap and randomization can be regarded as special cases of **Monte Carlo** methods.
- ▶ In the general **Monte Carlo** setting we decide a theoretical model that can be used to generate the set of measurements under the null hypothesis (instead of sampling from the set of observed data points)
- ▶ We can estimate the required probability by comparing the observed data with randomly generated data (like in the case of randomization)
- ▶ Typically, **Monte Carlo** tests are designed for a very specific test.