Lab 4: Like a Jackknife to the Heart

36-350, Statistical Computing

Friday, 21 September 2012

Agenda: Writing functions to automate repetitive tasks; writing functions that call other functions; quantifying uncertainty.

Last time, you fit a gamma distribution to data on the weights of cat's hearts. However, if we repeated the same measurements with a different sample of cats from the same population, we would almost certainly not get the same estimated parameters. How uncertain should we be about our estimates because of sampling noise?

There are many ways of quantify this sort of uncertainty, using more or less probability theory. This lab will introduce a computational way of estimating standard errors, called the **jackknife** which is fairly straightforward, though not always very accurate[1]. Here is how it goes:

- Get a set of n data points and get an estimate $\hat{\theta}$ for parameter of interest θ.
- For each data point i, remove i from the data set, and get an estimate $\hat{\theta}_{(-i)}$ from the remaining $n-1$ data points.
- Find the mean $\overline{\theta}$ of the n values of $\hat{\theta}_{(-i)}$
- The jackknife variance of $\hat{\theta}$ is

  $$\frac{n-1}{n}\sum_{i=1}^{n}(\hat{\theta}_{(-i)} - \overline{\theta})^2$$

  which works out to $(n-1)^2/n$ times the sample variance of the $\hat{\theta}_{(-i)}$ values[2].
- The jackknife standard error of $\hat{\theta}$ is the square root of the jackknife variance.

[1] Nobody seems to know where the name comes from, but it goes back at least to the 1950s. If you take advanced data analysis (36-402), you will learn a more accurate, but more complicated, descendant of the jackknife called the **bootstrap**.

[2] Can you explain why that's the right inflation factor? It may help to consider first the situation where n is large, so $(n-1)^2/n \approx n$.

The accompanying R file, `http://www.stat.cmu.edu/~cshalizi/statcomp/labs/04/lab-04.R`, contains a specimen `gamma.est` function, which you can use in place of whatever you wrote last time (or not)

1. (20) Take the first three cats from the cats from the data frame. Using `gamma.est`, estimate a and s for each of the three pairs of cats. Calculate the jackknife standard errors for a and s this would imply. Do not write a function for this; instead, repeatedly call `gamma.est`, store the results, and calculate variances from the estimates.

2. (25) Write a function, `gamma.jackknife()`, which takes a data vector and returns the jackknife standard errors in estimates of a and s. (It does not have to return the estimate itself.) It should call `gamma.est`. You may use a `for` loop.

3. (10) When run on the first three cats, does the output of `gamma.jackknife` agree with your calculation in problem 1?

4. (5) Use `gamma.jackknife` to find standard errors for a and s for the whole data.

5. (10) Estimate the a and s parameters separately for male and female cats. Find jackknife standard errors for both parameters for both estimates — four standard errors in all.

6. (15) When two independent estimated quantities d_1 and d_2 have standard errors s_1 and s_2, the standard error of their difference, $d_1 - d_2$, is $\sqrt{s_1^2 + s_2^2}$. Calculate the difference in both estimated parameters between the female and male cats, and the standard error of the differences.

7. (15) Does your answer to the previous question indicate that there is a significant sex difference in the distribution of cats' hearts' weights? Explain.