Lab 5

Statistical Computing, 36-350

Friday October 2, 2015

Today's agenda: Writing functions to automate repetitive tasks; fitting statistical models.

General instructions for labs. Upload an R Markdown file, named .Rmd", to Blackboard. You will give the commands to answer each question in its own code block, which will also produce plots that will be automatically embedded in the output file. Each answer must be supported by written statements as well as any code used. Include the name of your lab partner at the top of the file.

R Markdown setup. Open a new R Markdown file; set the output to HTML mode and click "Knit HTML". This should produce a web page with the knitting procedure executing your code blocks. You can edit this new file to produce your homework submission. Alternatively, you can start from the lab's R Markdown file posted on the course website, as a template.

The *gamma* distributions are a family of probability distributions defined by the density functions,

$$f(x) = \frac{x^{a-1}e^{-x/s}}{s^a\Gamma(a)}$$

where the gamma function $\Gamma(a) = \int_0^\infty u^{a-1}e^{-u}du$ is chosen so that the total probability of all non-negative x is 1. The parameter a is called the *shape*, and s is the *scale*. The gamma probability density function is called `dgamma()` in R. You can prove (as a calculus exercise) that the expected value of this distribution is as, and the variance as^2.
If the mean and variance are known, μ and σ^2, then we can solve for the parameters,

$$a = \frac{a^2s^2}{as^2} = \frac{\mu^2}{\sigma^2},$$
$$s = \frac{as^2}{as} = \frac{\sigma^2}{\mu}.$$

In this lab, you will fit a gamma distribution to data, and estimate the uncertainty in the fit. Our data today are measurements of the weight of the hearts of 144 cats. The data is contained in a data frame called `cats`, in the R package `MASS`. (This package is part of the standard R installation.) This records the sex of each cat, its weight in kilograms, and the weight of its heart in grams. Load the data as follows:

```
library(MASS)
data(cats)
```

Part I: Feline Hearts

1. Run `summary(cats)` and explain the results.
2. Plot a histogram of these weights using the `probability=TRUE` option. Add a vertical line with your calculated mean, using `abline(v=yourmeanvaluehere)`. Does this calculated mean look correct?
3. Calculate the mean, standard deviation, and variance of the heart weights using R's existing functions for these tasks. Plug the mean and variance of the cats' heart weights into the formulas above to get estimates (guesses of the true value, derived from your sample on hand) of a and s. What are they? Do not report them to more significant digits than is reasonable.

4. Write a function, `summary.stats()`, which takes as input a vector of numbers and returns the mean and variances of these numbers, in the form of a list. Confirm that you are returning the values from above. (You can use the existing mean and variance functions within this function.)

Part II: Gamma Hearts

5. Now, create a new function `gamma.estimates()` which again takes a vector of numbers, calculates the mean and variances using `summary.stats()`, and then returns the estimates of a and s, in a gamma distribution model. What estimates does it give on the cats' heart weights? Does it agree with your previous calculation?

6. Estimate the a and s separately for all the male cats and all the female cats, using `gamma.estimates()`. Give the commands you used and the results.

7. Now, produce a histogram for the female cats. On top of this plot, overlay the shape of the gamma PDF on the same plot in red. Is this distribution roughly consistent with the histogram? (Hint: you may use `curve()` with its first argument as `dgamma()`, the known PDF for the Gamma distribution. You may also use `lines()` to plot the values explicitly, using the PDF evaluated at some equally spaced weight values in the data range—e.g., `seq(from=somevalue,to=someothervalue,by=somesmallvalue)`.)

8. Repeat the previous step for male cats. How do the distributions compare? For a fair comparison, use the same x axis limits in the two histograms produced in questions 7 and 8.

9. Using the`rgamma()` function, generate a gamma distributed sample of the same size as the male cat heart weights, from their parameter estimates obtained in problem 8. Plot a histogram of this, side-by-side with a histogram of the male cats' original heart weights. For ease of comparison, use a fixed number (10) of breaks, using the option `breaks=10`. Do they look comparable?

10. Make a function that takes the vector of heart weights as input, and produces the same plots in the last problem. Test that it works, on both male and female cats. (Specifically, it should do all of the necessary calculations *within* the function, and shouldn't use globally declared variables, for full credit. It does not need to have specific x limits.)

11. **Bonus (1pt)**: Modify the above function to take as an additional input `xlim` that sets the x axis limits for the two plots it creates, and with a default value of `c(0,5)`. This allows for a fairer comparison by eye, of the original sample and the generated samples. Use it to plot the same plots as in problem 10, optionally with x axis limits that *you* think sensibly shows everything in the data range, but doesn't overdo it.