7 · *More probability and hypothesis testing*

Due Monday, March 9, 2015

(1) Probability mass functions and maximum likelihood

Suppose that you point your telescope at a distant, faint light source and track the number of photons that arrive at your detector during each one-second interval. In the first one-second interval, you observe x_1 photons; in the next one-second interval, you observe x_2 photons; and so on.

The question at issue is: how should you estimate the mean rate of photon arrival? That is, you want λ: the expected number of photons that will arrive in any given one-second interval.

(A) Give at least two reasons why the assumption of normally distributed photon counts, $X \sim \mathrm{N}(\lambda, \sigma^2)$, is not ideal here. (Note: we use big X to describe the random variable, and little x_i to denote the observed value in the ith period.)

(B) In light of (A), you decide to model the photon arrival counts with a Poisson distribution having rate parameter λ. This distribution has probability mass function

$$P(X = x \mid \lambda) = \frac{\lambda^x}{x!} \, e^{-\lambda} \, .$$

Derive an expression for the maximum-likelihood estimator for λ in terms of your n observations $x_1, \ldots, x_n$, assuming that each observation is independent. Show your work.

Two hints, applicable both here and elsewhere:

1. A function $f(x)$ will obtain a maximum at the same value of x where $\log f(x)$ obtains a maximum.
2. The product rule of differentiation is really annoying. But taking the log of a product turns it into a sum. . . .

(C) Suppose you observed 30 counts x_i, each corresponding to a one-second interval. These counts were:

```
12 18 15  8 17 13 22 13 13 13 12 11 15 15 12
 8 20 12 14 11  9 15 16 20  9 15 13 19 18 14
```

Using techniques you have learned, construct an approximate 95% confidence interval for λ using the maximum-likelihood estimator.[1]

[1] You can scan a raw text file into R using the "scan" function.

(2) Testing in regression

Do smarter animals dream more, all else being equal? To investigate, return to the "mammalsleep" data set from a previous homework (in the "faraway" library). As a preliminary step, construct a proxy measure of higher-level brain activity by regressing log brain weight upon log body weight, and extracting the residuals from this regression. Let's call this new variable "smarts" and treat it as a proxy for the intelligence of a species. You should compare these residuals to the actual animals, perhaps with the R command `cbind(mammalsleep, smarts)`, to ensure that this proxy looks sufficiently sensible to pass the sniff test.

(A) Plot dreaming hours (Y) versus smarts (X). Fit a simple linear model and summarize your findings from this simple analysis.

(B) An obvious potential confounder is the number of hours per night an animal sleeps. Thus the goal is to estimate the marginal effect of a change in residual brain size (smarts) upon nightly dreaming hours, adjusting for this confounder. Fit a multiple regression model that correctly estimates the marginal effect of smarts on dreaming hours, holding sleeping hours constant. Report your conclusions and note any shortcomings you perceive in the model.[2] Use an appropriate permutation test in conjunction with this model to assess the evidence against the null hypothesis that smarts and dreaming hours are unrelated, once sleeping hours have been accounted for.

[2] All good scientific papers comment honestly on their own limitations.