Lab 7: Bunches of Novels

36-350

25 October 2013

AGENDA: Working with simulations; transforming data between different representations; calculating sampling distributions by simulation.

The file `http://www.stat.cmu.edu/~cshalizi/statcomp/13/labs/07/moretti.csv` contains data compiled by the literary scholar Franco Moretti on the history of genres of novels in Britain between 1740 and 1900 (Gothic romances, mystery stories, stories, science fiction, etc.). Each record shows the name of the genre, the year it first appeared, and the year it died out.

It has been conjectured that that genres tend to appear together in bursts, bunches or clusters. We want to know if this is right. We will simulate what we would expect to see if genres really did appear randomly, at a constant rate — a **Poisson process**. Under the assumption, the number of genres which appear in a given year should follow a Poisson distribution with some mean λ, and every year should be independent of every other.

1. (10) If Poisson variables $x_1, x_2, \ldots x_n$ are independent and all have the same mean λ, the likelihood function is

   $$L(\lambda) = \prod_{t=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}$$

   Write a function, `poisLoglik`, which takes as inputs a single number `lambda` and a vector `data`, and returns the *log*-likelihood of that parameter value on that data. What should the value be when `data = c(1,0,0,1,1)` and `lambda=1`? Why do you get weird results when `lambda=0`? *Hint:* `dpois`.

2. (10) Write a function, `count_new_genres`, which takes in a year, and returns the number of new genres which appeared in that year: 0 if there were no new genres that year, 1 if there was one, 3 if there were three, etc. What should the values be for 1803 and 1850? Does your function work properly for those years?

3. (20) Create a vector, `new_genres`, which counts the number of new genres which appeared in each year of the data, from 1740 to 1900. What positions in the vector correspond to the years 1803 and 1850? What should those values be? Is that what `new_genres` has?

4. (5) Plot `poisLoglik` as a function of λ on the `new_genres` data. (If the maximum is not at $\lambda = 0.273$, you're doing something wrong.)

5. (10) To investigate whether genres appear in bunches or randomly, we look at the spacing between genre births. Create a vector, `intergenre_intervals`, which shows how many years elapsed between new genres appearing. (Explain how you handle years with multiple new genres.) What is the mean of the time intervals between genre appearances? The standard deviation? The ratio of the standard deviation to the mean, called the **coefficient of variation**? *Hint:* `diff`.

6. (40) For a Poisson process, the coefficient of variation is expected to be around 1. However, that calculation doesn't account for the way Moretti's dates are rounded to the nearest year, or tell us how much the coefficient of variation might fluctuate. We will handle both of these by simulation.

   (a) (5) What command do you use to generate a vector of n independent Poisson variables, each with mean λ?

   (b) (20) Write a function which takes a vector of numbers, representing how many new genres appear in each year, and returns the vector of the intervals between appearances. Check that your function works by seeing that when it is given `new_genres`, it returns `intergenre_intervals`.

   (c) (15) Write a function to simulate a Poisson process and calculate the coefficient of variation of its inter-appearance intervals. It should take as arguments the number of years to simulate and the mean number of genres per year. It should return a list, one component of which is the vector of inter-appearance intervals, and the other their coefficient of variation. Run it with 141 years and a mean of 0.273; the mean of the intervals should generally be between 3 and 4.

7. (5) Run your simulation 100,000 times, taking the coefficient of variation (only) from each. (This should take less than two minutes to run.) What fraction of simulation runs have a higher coefficient of variation than Moretti's data?

8. (10) Explain what this does and does not tell you about the conjecture that genres tend to appear together in bursts.