STAT GU4206/GR5206 Homework 7 [100 pts]
Due 11:59pm Friday, December 1 on Canvas

Your homework should be submitted on Canvas as an `R` Markdown file. **Please submit the knitted .pdf or .html file** along with the .Rmd file. We will not accept any other formats. Please clearly label the questions in your responses and support your answers by textual explanations and the code you use to produce the result. Note that you cannot answer the questions by observing the data in the "Environment" section of `R`Studio or in Excel – you must use coded commands. Please do not waste space by printing the dataset or any vector over, say, length 20.

Goals: More practice with simulations. Summarizing data using distributions and estimating parameters.

The file `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.

i. Assume the variables $x_1, x_2, \ldots, x_n$ are independent and Poisson-distributed with mean λ then the log likelihood function is given by the following:

$$\ell(\lambda) = \sum_{i=1}^{n} \log\left(\frac{\lambda^{x_i} e^{-\lambda}}{(x_i)!}\right).$$

Write a function `poisLoglik`, which takes as inputs a single number λ 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 $\lambda = 1$?

ii. 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?

iii. 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 your vector `new_ genres` has for those years?

iv. 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.)

v. Use `nlm()` to maximize the log likelihood to check the $\lambda = 0.273$ value suggested in the previous question. Hint: you may need to rewrite your function from (i.) with some slight alterations.

vi. 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. (If two genres appear in the same year, there should be a 0 in your vector, if three genres appear in the same year your vector should have two zeros, and so on. For example if the years that new genres appear are $1835, 1837, 1838, 1838, 1838$ your vector should be $2, 1, 0, 0$.) 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: The `diff()` function might help you here. Check out `?diff`.

vii. 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. 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`.

   b. 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 161 years and a mean of 0.273; the mean of the intervals should generally be between 3 and 4.

viii. Run your simulation $10,000$ times, taking the coefficient of variation (only) from each. (This should take less than two minutes to run.) What fraction of simulations runs have a higher coefficient of variation than Moretti's data?

ix. Explain what this does and does not tell you about the conjecture that genres tend to appear together in burst?