STAT GR5206 Homework 4 [100 pts]
Due 8:00pm Monday, November 27 on Canvas

Your homework should be submitted on Canvas using RMarkdown. Please submit both a knitted .pdf file and a raw .Rmd file. (If you are having trouble knitting to .pdf come to office hours and we'll try to sort it out, but for the homework, knit to .html and then convert to .pdf before handing it in). We will not (and cannot) 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 RStudio or in Excel – you must use coded commands.

Goals: Practice with simulating distributions via the Inverse Transform Method. Summarizing data using distributions and estimating parameters.

Part 1

We continue working with the World Top Incomes Database, and the Pareto distribution, as in previous labs and homework. Recall that for most countries in most time periods, the upper end of the income distribution roughly follows a Pareto distribution, with probability density function

$$f(x) = \frac{(a-1)}{x_{min}} \left(\frac{x}{x_{min}} \right)^{-a} \tag{1}$$

for incomes $x \geq x_{min}$. In a previous homework, we estimated the parameter a based on the `wtid-report` dataset for years ranging from 1913 to 2015.

Now suppose that we are interested in simulating the upper end of income just for 2015 using the Pareto distribution, (1). Let the 'upper end' begin at the 99th annual income percentile for 2015 (meaning, we let $x_{min} = \$407,760$) and we'll estimate the Pareto exponent using $\hat{a} = 2.654$ which we calculated in the previous homework. Then we can model the upper end of income for 2015 by the Pareto distribution having pdf

$$f(x) = \frac{\hat{a}-1}{x_{min}} \left(\frac{x}{x_{min}} \right)^{-\hat{a}} = \frac{1.654}{407760} \left(\frac{x}{407760} \right)^{-2.654} \tag{2}$$

Perform the following tasks:

i. Define a function `f` which takes three inputs x, a vector, and scalars a and x_{min} having default values of $a = \hat{a}$ and $x_{min} = \$407,760$. The function should output $f(x)$ for a given input $x > x_{min}$. Plot the function between x_{min} and $1,000,000$. Make sure your plot is labeled appropriately.

ii. For $x > x_{min}$, the cdf equals

$$F(x) = 1 - \left(\frac{x}{x_{min}}\right)^{-a+1}$$

Find the inverse function $F^{-1}(u)$ and define a function `upper.income`. The function should have three inputs u, a vector, and scalars a and x_{min} taking default values of $a = \hat{a}$ and $x_{min} = \$407,760$. The function should output $F^{-1}(u)$ for a given input $u \in (0,1)$. Make sure `upper.income(.5)` returns 620020.2.

iii. Using the Inverse Transform Method, simulate 1000 draws from the Pareto distribution (1) and plot a histogram of your values. Overlay the simulated distribution with the Pareto density (1). Make sure to label the histogram appropriately.

iv. Using your simulated set, estimate the median income for the richest 1% of the world. Recall from lab that the proportion of people whose income is at least x_{min} whose income is also at or above any level $w \geq x_{min}$ is

$$\mathbf{Pr}(X \geq w) = \left(\frac{w}{x_{min}}\right)^{-a+1}.$$

Compare your estimated 50th percentile to the actual 50th percentile of the Pareto distribution.

Part 2

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. Return the value your function calculates 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. Return the value your function calculates 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 using the `new_ genres` data. I plotted λ ranging from 0 to 3 (note that $\lambda = 0$ has a log-likelihood of `-Inf`, but it shouldn't cause problems). The maximum should be $\lambda = 0.273$.

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 may be helpful. Check out `?diff` and run `diff(c(1835, 1837, 1838, 1838, 1838))`.

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 simulating data from a Poisson process with the same mean number of new genres per year and then calculating for the simulated data an `intergenre_intervals` vector on which we can calculate the coefficient of variation. We'll see how often our simulated data produces a coefficient of variation value as or more extreme than the one form our data.

   a. Write a line of code that generates 161 random draws from a Poisson distribution with $\lambda = 0.273$.

   b. Write function that takes as input a vector of numbers representing how many new genres appear in each year (e.g. the output of the code you wrote in the previous part of the question) and returns the vector of the intervals between appearances. Check that the function works by seeing that when it is given `new_genres`, it returns `intergenre_intervals`. I did this by using the input to create a vector corresponding to the years in which new genres began (like `genres$Begin`) and then used my code from the previous question, but you can do anything that works.

   c. 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 called `num.years` and the mean number of genres per year called `mean.genres`. It should return a list, one component of which is the vector of inter-appearance intervals of the simulated data, and the other the coefficient of variation of the data. Note that to produce the desired output you will need to use the previous parts of this question. Run it with 161 years and a mean of 0.273.

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?

The above result tells us that there isn't really any evidence that new genres appear together in bursts. If the appearance of new genres were truly random (not clustered), meaning they appear according to a Poisson process, then around 23 precent of the time we would see results as or more clustered than Moretti's data (when we consider the coefficient of variation as a measure of the amount of cluster). Interestingly, more subtle tests than the one used above indicate that there is some evidence of bursts of genre formation.