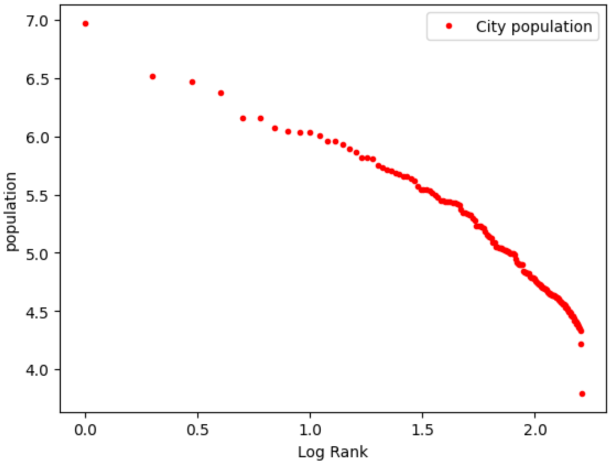
Mathematical Data Science HW6

20180617 You SeungWoo

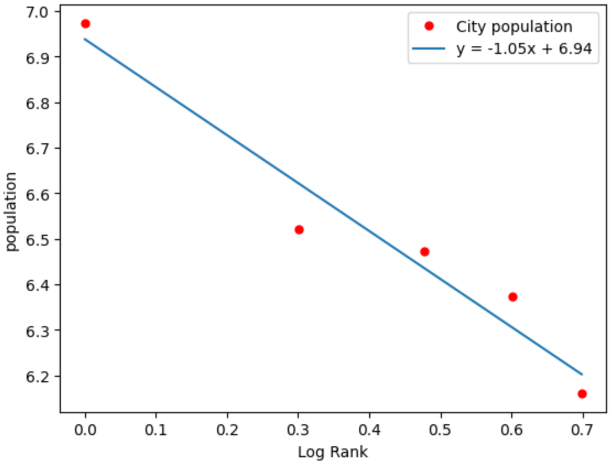
November 9, 2023

Problem 1

Solution. My city is Yeongju which is ranked at 72 with 108,443. The below is a population distribution of all cities in Korea based on log scale.



To determine the given constants, use the linear regression to minimize the squared error.



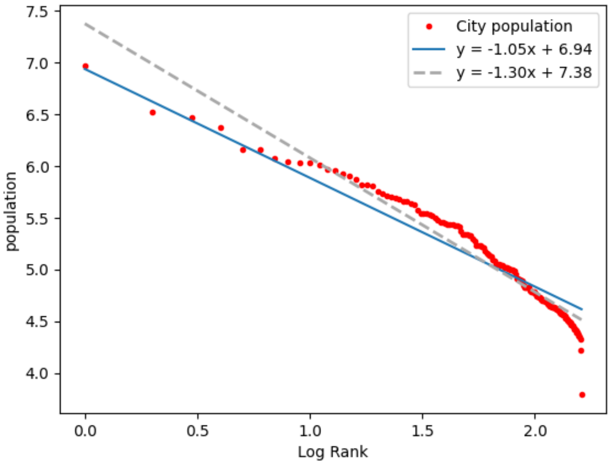
Note that I just put $C = 0$. Even though we can change C, it is impossible to find a value that perfectly fits the five pieces of data. By the log scale transformation, we have:

$$\log P(r) = \beta \log (r + C) + \log \alpha$$

From the above line, we get

$$\beta = -1.05$$
$$\log \alpha = 6.94 \Rightarrow \alpha = 10^{6.94}$$

By this result, we have the relation between total cities and the line.

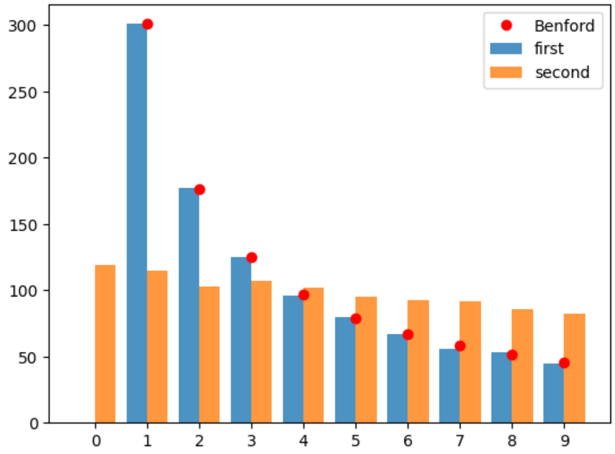


The line $y = -1.30x + 7.38$ is the result of regression using all cities. They are mostly linearly distributed, but we can see that regression does not work well for some data. Clear outliers at both ends prevent Zipf's law from working. Still, excluding these things, urban distribution can be explained as almost following Zipf's law. There are a variety of variables that may be involved in this alignment. If there is a primate city, like Seoul, it can deviate from this rule depending on how much control it has. There may be significant differences in population depending on the level of economic and industrial growth in the region. In the case of Ulleung-gun, which is ranked last, it can be seen that the population is extremely low due to the geographical feature of being a city made up of very small islands.

□

Problem 2

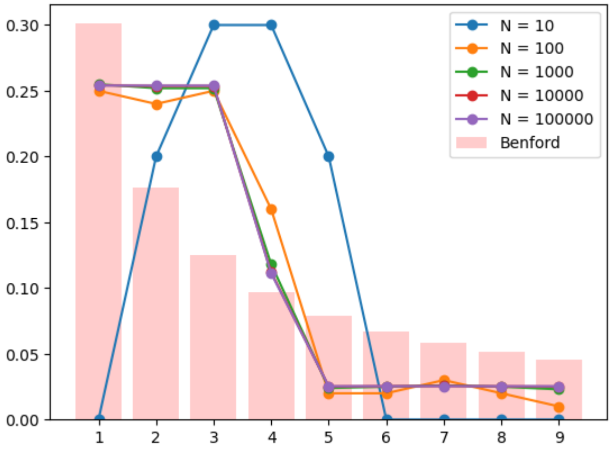
Solution. Here is the result. We can see that the second digits do not follow the Benford's law. Note that the first few numbers of the Fibonacci sequence are single digits, so they are excluded from statistics of `second`.



□

Problem 3

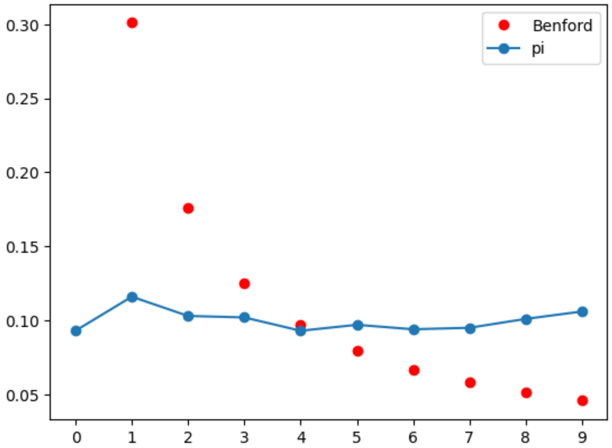
Solution. Here is the result. We can see that the primes do not follow the Benford's law.



□

Problem 4

Solutions. Here is the result. We can see that the numerical digits of π do not follow the Benford's law.



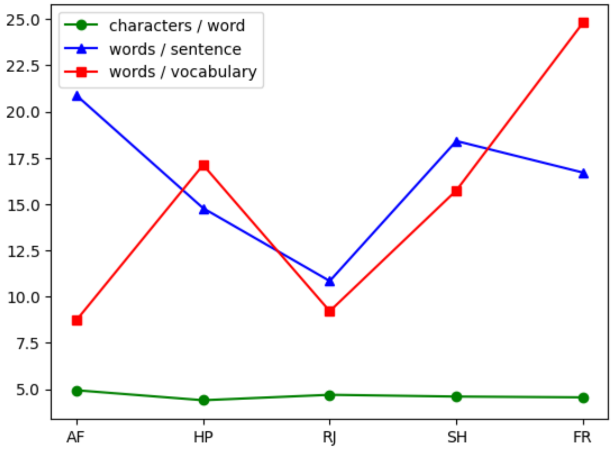
□

Problem 5

Solution. I choose the following list:

- Animal Farm by George Orwell
- Harry Potter and the Sorcerer's Stone by Joanne Kathleen Rowling
- Romeo and Juliet by William Shakespeare
- The Adventures of Sherlock Holmes by Arthur Conan Doyle
- The Fellowship of the Ring by John Ronald Reuel Tolkien

Here is the result. We can see that the ratio analysis is sufficiently effective for the comparision.



For this comparision, we can think about:

- Vocabulary and Grammar: Specific word choice, complexity of sentence structure, and use of sentence types
- Point of view and narrator: First-person and third-person, differences in narrative structure
- Literary techniques: various literary expression methods including metaphor, irony, etc.
- Morphemes: Frequency of use in words containing adverbs and adjectives

□

Problem 6

Solution. 저자는 빅데이터가 지닌 문제를 어떻게 스몰데이터로 극복할 수 있는지 설명한다. 빅데이터는 상관관계 파악에 유리하지만, 빅데이터 자체로는 통찰력을 이끌어낼 수 없으며 분석 중심으로 이루어져 있다. 그렇기에 빅데이터만으로는 개개인의 성향이나 기호에 맞춘 분석이 어려울 수 있다. 하지만 스몰데이터를 이용하면 이를 효과적으로 파악할 수 있다. 따라서 스몰데이터와 빅데이터의 장점을 모두 활용해야 진정으로 상황에 맞게 적절한 분석이 가능하다는 입장이다.

실제로 빅데이터 기반의 공부를 할 때, 빅데이터가 가진 장점에 대해서만 배웠던 기억이 있다. 우리는 기존의 감정적인 결정에서 벗어나, 빅데이터와 분석에 기반한 결정을 내릴 수 있어야 한다고 설명한다. 또한 빅데이터는 작은 변동을 무시하고 전체의 변화를 파악한다. 수 많은 데이터가 넘쳐나는 현대사회에는 빅데이터 시대가 불가피하다고 생각한다. 그래도 저자의 주장을 통해 그럼에도 스몰데이터만이 할 수 있는 일이 있다고 느꼈다. 빅데이터가 하지 못하는, 수 많은 데이터의 경향에 벗어나는 값을 세세하게 분석하는 등의 방법으로 스몰데이터를 활용할 수 있을 것이다.

□