Assignment

A set of numbers satisfies Bernford's law if the leading digit d occurs with probability

$$\begin{aligned}\pi_d &= \log_{10}(d+1) - \log_{10}(d) \\ &= \log_{10}\left(\frac{d+1}{d}\right) \\ &= \log_{10}\left(1+\frac{1}{d}\right)\end{aligned}$$

Therefore, the frequency distribution of the leading digit d in such a set is the following: $\pi_1 = 30.1\%$, $\pi_2 = 17.6\%$, $\pi_3 = 12.5\%$, $\pi_4 = 9.7\%$, $\pi_5) = 7.9\%$, π_6, $\pi_7 = 5.8\%$, $\pi_8 = 5.1\%$ and $\pi_9 = 4.6\%$.

In this assignment, you will examine this law when applied to a large online retail dataset. You have price data in "online_retail.csv" for about 500,000 items from different countries. You will ignore all items that start with 0.

You compute the real distribution $F = (f_1, f_2, \ldots, f_9)$ of frequencies of the leading digit in these prices and compare F to two models:

1. Model 1: equal-weight distribution: each leading digit has the same frequency $1/9 = 11.1\%$. In other words, your

predicted model of frequencies in this model is a 9-digit vector $P = (1/9, 1/9, \ldots, 1/9)$

2. Model 2: leading digit follows the Bernford's law. In this model, the prediction is a 9-digit vector $\pi = (\pi_1, \pi_2, \ldots, \pi_9)$

Questions:

1. plot 3 histograms for the frequencies for real distribution, equal-weight and Bernford (for each digit)
2. plot 3 histograms for the relative errors for Models 1 and 2 (for each digit)
3. compute RMSE (root mean squared error) for model 1, 2. Whick model is closer to the real distribution?
4. take 3 countires of your choice: one from Asia, one from Europe and one from the Middle East. For each of these countries do the following:
   (a) compute F, P and π
   (b) using RMSE as a "distance" metric, for which of these chosen three countries is the distribution "closest" to equal weight P?
5. discuss your findings