

# 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:  $\pi_1 = 30.1\%$ ,  $\pi_2 = 17.6\%$ ,  $\pi_3 = 12.5\%$ ,  $\pi_4 = 9.7\%$ ,  $\pi_5 = 7.9\%$ ,  $\pi_6 = 6.7\%$ ,  $\pi_7 = 5.8\%$ ,  $\pi_8 = 5.1\%$  and  $\pi_9 = 4.6\%$ .

In this assignment, you will examine this law when applied to product data from Nordstrom contained in "nordstrom\_product\_data.csv" from Kaggle.

<https://www.kaggle.com/datasets/polartech/300k-nordstrom-product-data/>

This dataset contains more than 300,000 product data of the Clothing/Shoes/Handbags/Jewelry/Accessories/Maternity/Tech Accessories from the women department. Data attributes include price, review count, ratings, promotions, category info, brand etc.

We will focus on the "Price\_Retail" column (we will refer to this as price). You compute the real distribution  $F = (f_1, f_2, \dots, f_9)$  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  $1/9 = 11.11\%$ . In other words, your predicted model of frequencies in this model is a 9-digit vector  $P = (1/9, 1/9, \dots, 1/9)$
2. Model 2: leading digit follows the Bernford's law. In this model, the prediction is a 9-digit vector  $\pi = (\pi_1, \pi_2, \dots, \pi_9)$

## Questions:

1. download data in Pandas dataframe
2. how many categories are there?
3. which category has most sub-categories and which one has the fewest sub-categories
4. what is the median and what is the average price for products in each category
5. plot three histograms for the frequencies for real distribution, equal-weight and Bernford (for each digit)
6. plot two histograms for the relative errors for Models 1 and 2 (for each digit)
7. compute RMSE (root mean squared error) for model 1 and 2. Which model is closer to the real distribution?
8. take 3 categories of your choice For each of these categories do the following: (a) compute  $F$ ,  $P$  and  $\pi$ . (b) using RMSE as a "distance" metric, for which of these chosen three categories is the distribution "closest" to equal weight  $P$ ?
9. discuss your findings