Analysis of Nutrient Concentrations in Food Display Universal Behaviour

Rahul Oberoi
2021555

Saumil Lakra
2021097

1. Abstract

The authors of Nutrient Concentrations in Food Display Universal Behaviour analyzed the nutrient content data across the full US food supply. They found that the concentration of each nutrient follows a universal single-parameter scaling law that accurately captures the eight orders of magnitude in nutrient content variability. Specifically, the distribution of nutrient concentrations follows a log-normal distribution with a constant logarithmic standard deviation across all nutrients. We have further dived into the research that was started by them in this paper by trying to recreate the results obtained by them, however, on the latest FNDDS and BRENDA datasets.

2. Introduction

In the paper **Nutrient Concentrations in Food Display Universal Behaviour**, we analyzed FNDDS and BRENDA data to show that the concentration of each nutrient follows a universal single-parameter scaling law that accurately captures the eight orders of magnitude in nutrient content variability. We tried to show that the universality is rooted in the biochemical constraints obeyed by the metabolic pathways responsible for nutrient modulation, allowing us to confirm the empirically observed scaling law and to predict its variability in agreement with the data. We propose that the natural nutrient variability in food can be quantitatively formalized. The analysis can provide a quantitative way to understand the impact of food processing on nutrient balance and health effects.

Chemical concentrations in food are modulated by a densely wired biochemical reaction network, suggesting that the concentrations of individual components may follow familiar patterns, governing their expected values and the extent of their fluctuations across the food supply.

3. Dataset

3.1. Size and Shape of Dataset

The original dataset constituted the combination of FNDDS (2007-10) and FNDDS Flavonoids (2009-10). The dataset consisted of **7253** rows (which we were able to match) however it was further reduced down to the food consumed by individuals over the period of two days reported in the NHANES dataset. Food items consumed by babies were also removed from this dataset. This brought the number down to **4968** (4889 in the original paper) food items.

3.2. Plotted Figures

These figures have been plotted directly from the dataset obtained using as described in Section 3.1 and in Figure 1.

Figure2 shows the nutrients present in ”Raw Apple” on the y-axis and the x-axis represents the amount of nutrient.

Table2 shows the amount of Zinc and Thiamin present (in grams) in the five food items.

Feature	**Description**	**Type of Feature**
Food Index	Unique identifier of a particular food item	Integer
Nutrient Index	Unique identifier of a particular nutrient	Integer
From	Starting date from which the food item was sampled	Date
To	Ending date till the food item was sampled	Date
Amount	Amount of nutrient present in the food item	Float
Nutrient	Name of the nutrient	String
Unit	Unit in which the amount of nutrient has been reported	String
Food	Name of the food item	String
Nutrient Category	Category of the nutrient (present only in flavonoids)	String

Table 1. Filtered Dataset Overview

Food	**Zinc (in grams)**	**Thiamin (in grams)**
Oysters, raw	0.0393	0.000018
Apple, raw	0.00004	0.000017
Cherries, maraschino	0.00026	0
Yeast	0.00794	0.010990
Margarine, NFS	0	0.000007

Table 2. Concentration of Zinc and Thiamin

4. Methodology

4.1. Formulae

Given the wide range of food and drinks available to the consumer in grocery stores and restaurants and their home-cooked variants, a key determinant of nutrient exposure is the probability $P(x_n)$

$$P(x_n) = (1 - p_n)\delta(x_n) + p_n Q(x_n) \tag{1}$$

- $P(x_n)$: Probability that an individual or a population is exposed to x_n grams of nutrient n in a randomly consumed dish.
- p_n: Probability that a nutrient n is present in a random dish.
- $Q(x_n)$: Probability that the selected time carries x_n grams

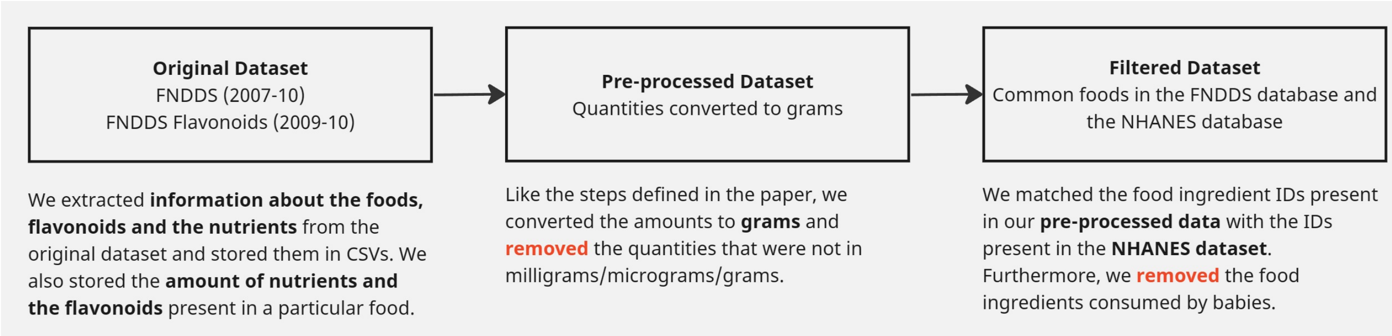


Figure 1. Procedure followed to reach the filtered dataset

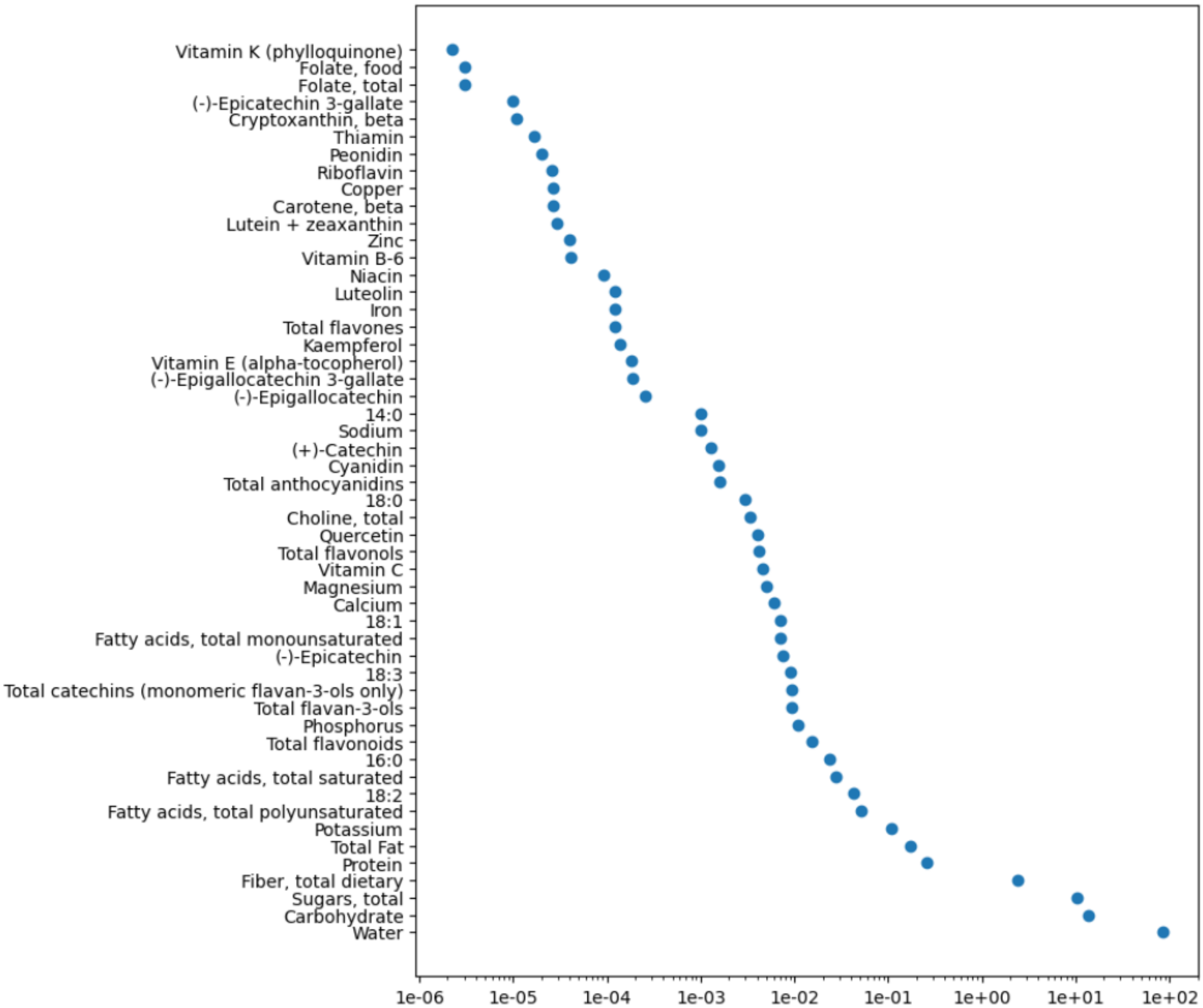


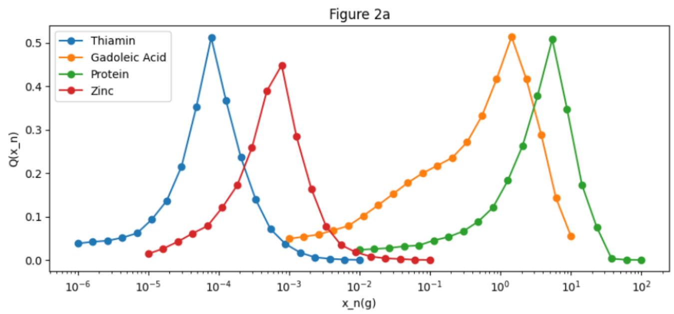
Figure 2. Concentration of nutrients in "Raw Apple"

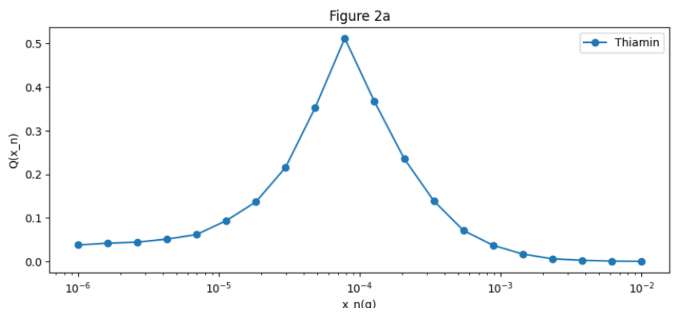
of nutrient n.

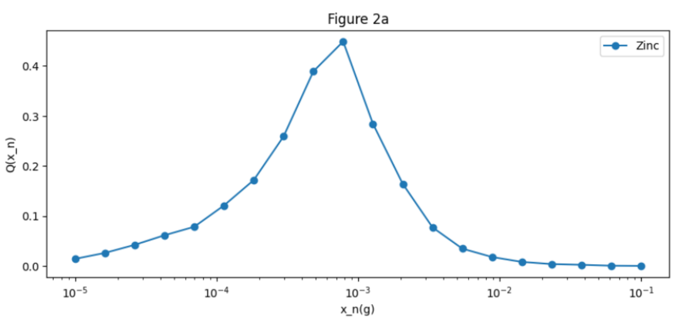
- $\delta(x_n)$: Dirac delta function evaluated at x_n

Figures 4.1 were plotted using the above equation.

The probability $Q(x_n)$ plays a fundamental role in nutrient exposure, capturing the food source variability of nutrient n available to the population. Indeed, individuals sample foods from the food supply according to their dietary pattern, and a precise description of $Q(x_n)$ is instrumental to quantify how nutrient intake varies in the population and the likelihood of observing extreme values and deficiencies.







4.1.1 Universal Scaling Law for Nutrient Content

To assess the diversity in nutrient levels across the food spectrum, we measured $Q(x_n)$ for 99 different nutrients found in 4,968 food items (according to the latest FNDDS data). $Q(x_n)$ is calculated for four nutrients, namely, Thiamin (a vitamin), Zinc (a mineral), Gadoleic acid (a fatty acid) and Protein, to capture their distribution across all the foods in the database.

4.1.2 Logarithmic Standard Deviation s_n

It is used to capture the variability of nutrient n across all foods, which appears to be the same for each of the four nutrients. It suggests that the degree of variability in nutrient content across all foods is independent of nutrient concentration. For these nutrients, s_n came out to be $1.51480.339$ and was calculated using the formula:

$$s_n = \sqrt{\langle(\log x_n)^2\rangle - \langle\log x_n\rangle^2}$$

4.1.3 Symmetry

According to Fig 2a of [1], $Q(x_n)$, is symmetrical. We calculated the logarithmic skewness of four nutrients.

Observing the patterns, we conclude that $Q(x_n)$ follows single distribution i.e. log-normal family and can be represented by the formula:

$$Q(x_n) = \frac{1}{x_n s_n \sqrt{2\pi}} e^{-\frac{(\log x_n - m_n)^2}{2s_n^2}} \tag{2}$$

- x_n is the weight of nutrient n in grams in a particular food
- s_n is the logarithmic standard deviation
- m_n is calculated using the formula:

$$m_n = \ln(\mu_n) - 2s_n^2$$

where $_n$ is the average concentration and s_n is the logarithmic standard deviation.

Linear average concentration is related to the linear standard deviation using the formula:

$$\sigma_n = \mu_n \times \sqrt{e^{s_n^2} - 1} \tag{3}$$

After encountering difficulties replicating the graphs of $Q(x_n)$, we reached out to one of the paper's authors, Giulia Menichetti. She advised us to remove the rows from the data where the concentration of nutrients equaled 0 grams. Following this adjustment, the graphs we plotted aligned with those presented in the paper.

4.1.4 Reverse Engineering

Since we were facing quite a few issues in replicating the results using equation 2, we decided to reverse engineer equation 1, i.e. we would pick a point from the figure and try to apply the reverse steps to it. For instance, in Figure 1c of [1], the value of ”Raw Apple” is 0.015 (approx) multiplying it with the probability of thiamin being present in a food item (0.9635), then on multiplying this with the number of food items in our dataset we get 718 which is nearly the number of food items that contain 0.0001 g or less of thiamin.[1]

5. Results

Using equation 1, 2 and 3 and plotting graphs: we find that nutrient concentrations in the food supply follow a single family of distributions that depend only on a single parameter, the average concentration $_n$ of nutrient n across all foods.

5.1. Linear Average Concentration(μ_n) vs Logarithmic Standard Deviation(s_n)

Linear average concentration is on the x-axis and is in grams, logarithmic standard deviation is on the y-axis (Figure 3).

5.2. Linear Average Concentration(μ_n) vs Skewness

Linear Average Concentration is on the x-axis and is in grams and skewness which is the measure of asymmetry in the distribution about its mean is on the y-axis (Figure 4).

[1]Note: It is our assumption that this is the correct way to interpret this equation (as confirmed with Dr. Ganesh Bagler).

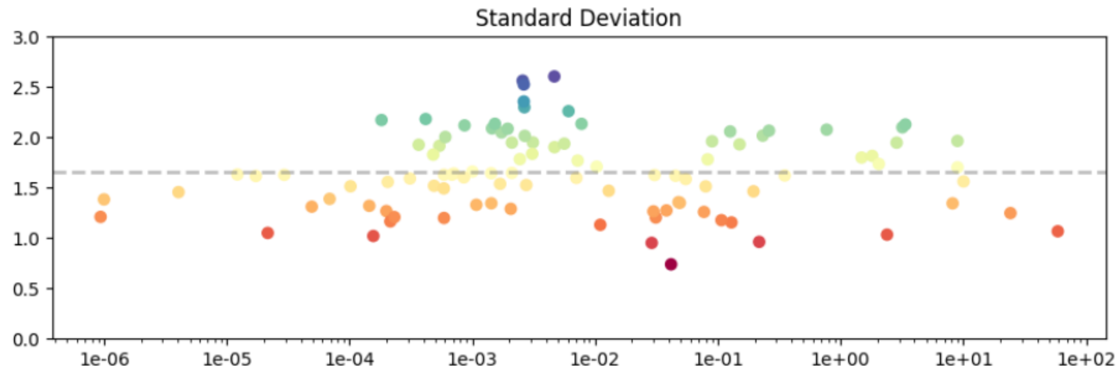


Figure 3.

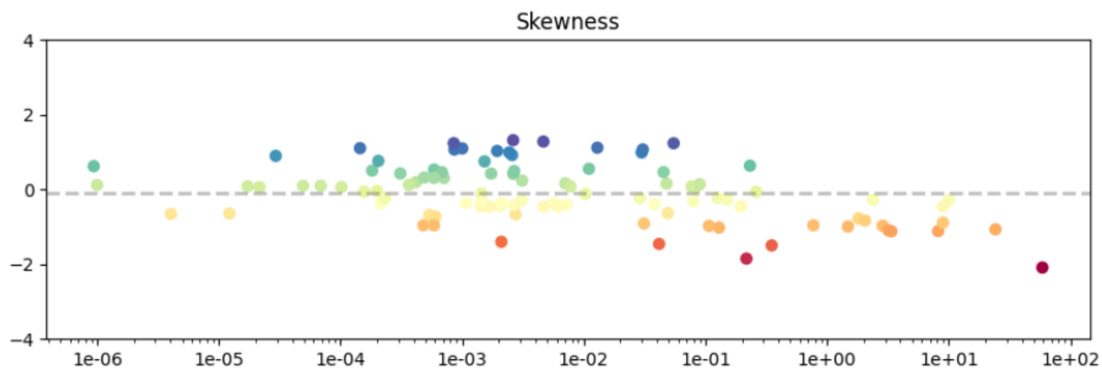


Figure 4.

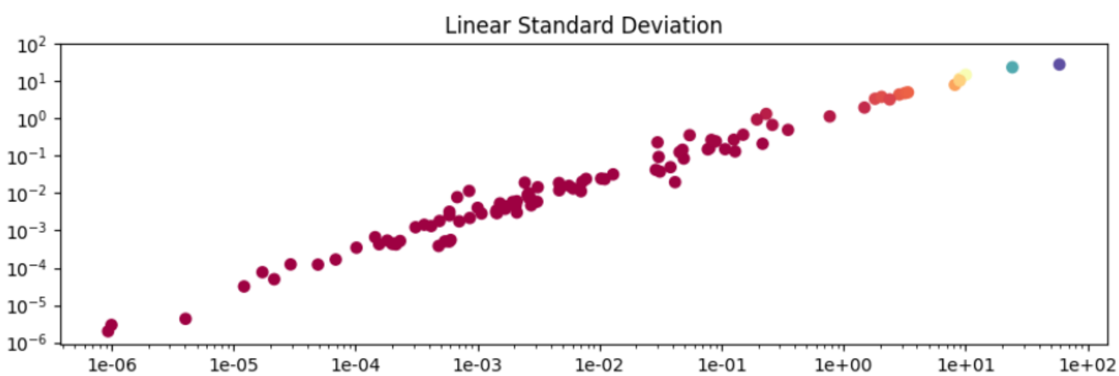


Figure 5.

5.3. Linear Average Concentration(μ_n) vs Linear Standard Deviation(σ_n)

Linear Average Concentration is on the x-axis and is in grams and Linear Standard Deviation is on the y-axis (Figure 5).

6. Future Work

BRENDA analysis is an attempt to achieve the same results that we got in food analysis and expand it to substrates that are the building blocks of food. Even though, we weren't able to reproduce the results using the BRENDA dataset due to computational and time constraints. Moreover, the dataset used originally in [1] is no longer available which made it even harder to match our results. Research can be continued further to match the results obtained using the FNDDS dataset on a substrate level.

References

[1] Giulia Menichetti and Albert-László Barabási. Nutrient concentrations in food display universal behaviour. *Nature Food*, 3(5):375–382, May 2022.