भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

EXPLORING BENFORD'S LAW FOR PRODUCTS AND SOME OTHER ARITHMETIC OPERATIONS

EP4130 Project - 2023

Vibhavasu Pasumarti*and Raghav Juyal†

EP20BTECH11015 and EP20BTECH11018
Engineering Physics
Indian Institute of Technology, Hyderabad

Instructor

Dr. Shantanu Desai

Department of Physics
Indian Institute of Technology, Hyderabad

Saturday 29th April, 2023

*ep20btech11015@iith.ac.in
†ep20btech11018@iith.ac.in

Contents

ABSTRACT

It has been observed that if $X_1, \ldots, X_M$ are independent continuous random variables with densities $f_1, \ldots, f_M$, as $M \to \infty$ the distribution of the digits of the product $X_1 \ldots X_M$ converges to Benford's law [1]. In this project we verify it and explore how random variables generated from various distributions behave when other arithmetic operations are applied on them repeatedly.

1 Introduction

Benford's law, also known as the law of anomalous numbers or the first-digit law, is an observation that in many numerical data-sets, the leading digit is likely to be small [2]. A set of numbers in base $B \geq 2$ is d are said to be following Benford's law when the probability of the leading digit of a number is written as

$$P(D_1 = d) = \log_B(d+1) - \log_B(d) = \log_B(1 + \frac{1}{d}), \text{ where } d \in 1, \ldots, B-1 \tag{1}$$

or more specifically in base 10 as

$$P(D_1 = d) = \log_{10}(d+1) - \log_{10}(d) = \log_{10}(1 + \frac{1}{d}), \text{ where } d \in 1, \ldots, 9 \tag{2}$$

The cumulative distribution function for base B is given as

$$F(d) = P(D_1 \leq d) = \log_B(d+1), \text{ where } d \in 1, \ldots, B-1 \tag{3}$$

Here D_1 denotes the leading digit. For example,

$$D_1(\sqrt{2}) = D_1(1.414\ldots) = 1$$
$$D_1(\pi^{-1}) = D_1(0.3183\ldots) = 3$$
$$D_1(e^{\pi}) = D_1(23.14\ldots) = 2$$

In a more complete form, Benford's law is a statement about the joint distribution of all the digits. For every natural number n,

$$P((D_1, D_2, \ldots, D_n) = (d_1, d_2, \ldots, d_n)) = \log_B\left(1 + \left(\sum_{j=1}^{n} B^{n-j} d_j\right)^{-1}\right), \text{ where } d_j \in 1, \ldots, B-1 \tag{4}$$

Here $D_2, D_3, \ldots$ represent the second digit, third digit and so on.

In many instances in literature [3, 4, 5, 6], it has been observed that the product of two random variables is closer to follow Benford's distribution than the input variables and as the number of terms increase, the expression seems to approach Benford's distribution. To understand the distribution of the digits of $X_1 \ldots X_M$ all we need to understand $\log_B |X_1 \ldots X_M| \mod 1$ [1]. This leads to the equivalent problem of studying sums of random variables modulo 1 which is ideally suited for Fourier analysis. This leads to a variant of the central limit theorem which in this case states that the sum of M independent random variables modulo 1 tends to the uniform distribution; which, by exponentiation, leads to Benford's law for the product [7].

2 Procedure

In this project we simulate the result for the case of products of random variables (generated by various distributions) and then try to see the distributions for some other arithmetic operations.

Our procedure involves:

1. Initialize j = 1
2. Generate 1000*j random variables.
3. Calculate the products of the random variables in groups of j
4. Find their first digit and add it to the count of each digit (here we assume base 10)
5. Compute the χ^2 statistic and the p value of the samples with respect to the ideal Benford's distribution
6. Plot the frequency of each digit and compare with Benford's distribution.
7. Increase j by 1 and repeat steps 2 - 6

We use a similar procedure to see the distribution of the first digits of random variables after repeated application of other arithmetic operations like addition, logarithm, exponents and sine. We also check the case for products by generating the random variables from different distributions.

Here we have used the χ^2 goodness of fit test [8, 9, 10, 11, 12] to check how likely it is for the distribution of the random variables to be the same as the Benford's distribution. In this test our null hypothesis is that the categorical data follows Benford's law. We define the χ^2 statistic of categorical data for a given expected frequency distribution as

$$\chi^2 = \sum \frac{(O-E)^2}{E} \tag{5}$$

Here O is the observed frequency and E is the expected frequency observations. After this we check a contingency table [13] which tells our p value which is the probability of a larger value of χ^2. If the p value is lesser than 0.05 we can reject the null hypothesis that the data follows the expected distribution. Else we cannot reject the null hypothesis and the data does not support the alternative hypothesis which is that the data follows a different distribution.

3 Observations and Analysis

In this section we show the plots we had observed. The code for generating these plots can be found at [14]

3.1 Product

In figure 1, we see the plot of the frequency of 1st digit of the products of j random variables taken at a time. We have generated the random variables using uniform, normal and log normal distributions. In all three cases we see that the products tend to follow Benford's law as j increases. We see that the p values are > 0.05 which implies that the null hypothesis (data follows Benford's law) cannot be rejected and the alternate hypothesis (data does not follow Benford's law) is not supported by the data. This is as we expected based on [1].

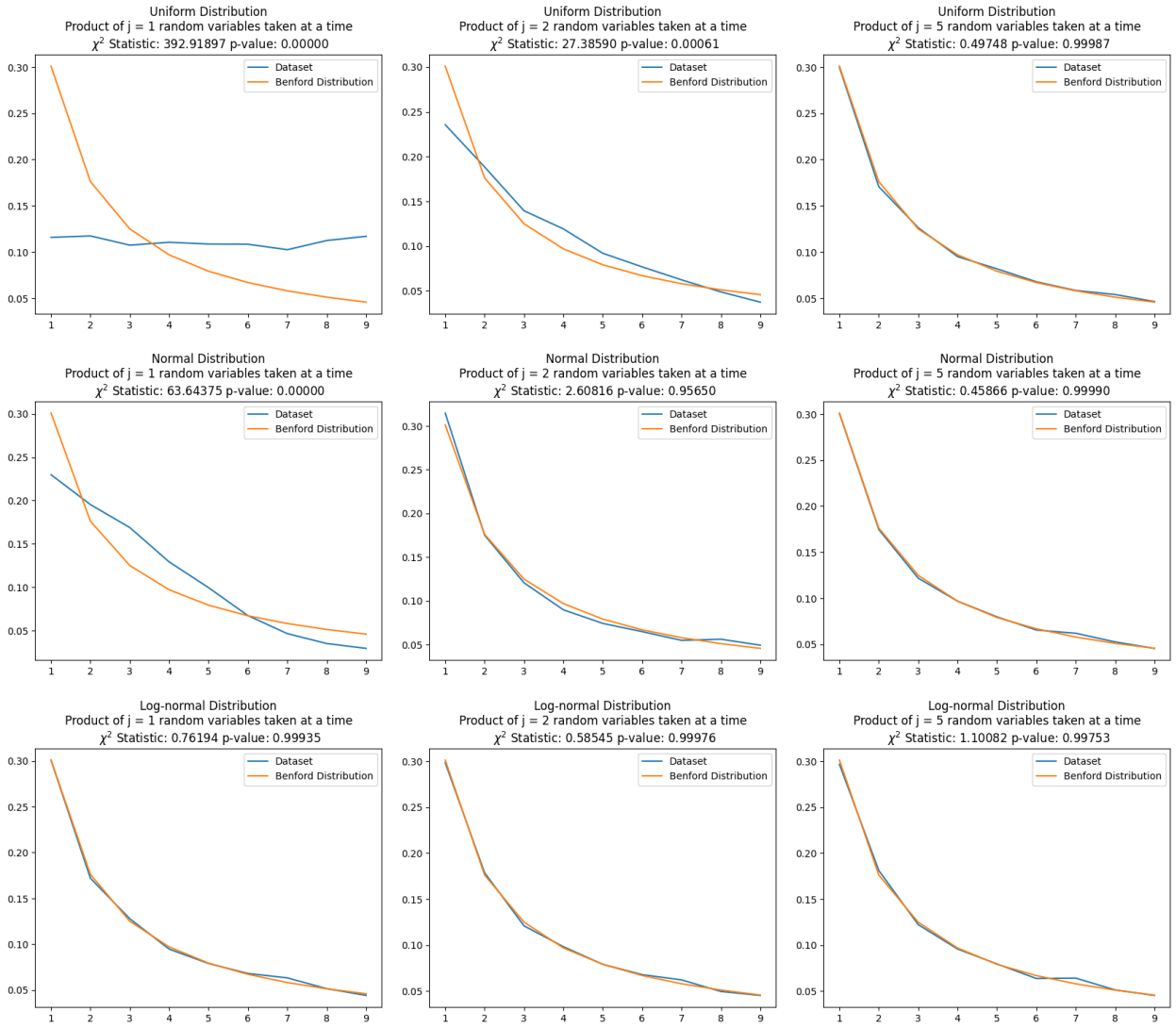

Figure 1: Taking products of j random variables at a time. The random variables are uniformly distributed in the 1st row, normally(Gaussian) distributed in the 2nd row and log normally distributed in the 3rd row.

3.2 Sum

In figure 2, we see the plot of the frequency of 1st digit of the sum of j random variables taken at a time. We have generated the random variables using uniform, normal and log normal distributions. In the first 2 cases we see that the sum does not tend to follow Benford's law as j increases. We see that the p values are < 0.05 which implies that the null hypothesis (data follows Benford's law) can be rejected and the alternate hypothesis (data does not follow Benford's law) is supported by the data. In the third case we see that the sum tends to follow the Benford's law as j increases and that $p > 0.05$. This is a special case since the sum of the log of numbers is the log of their products which we expect to follow Benford's law [1].

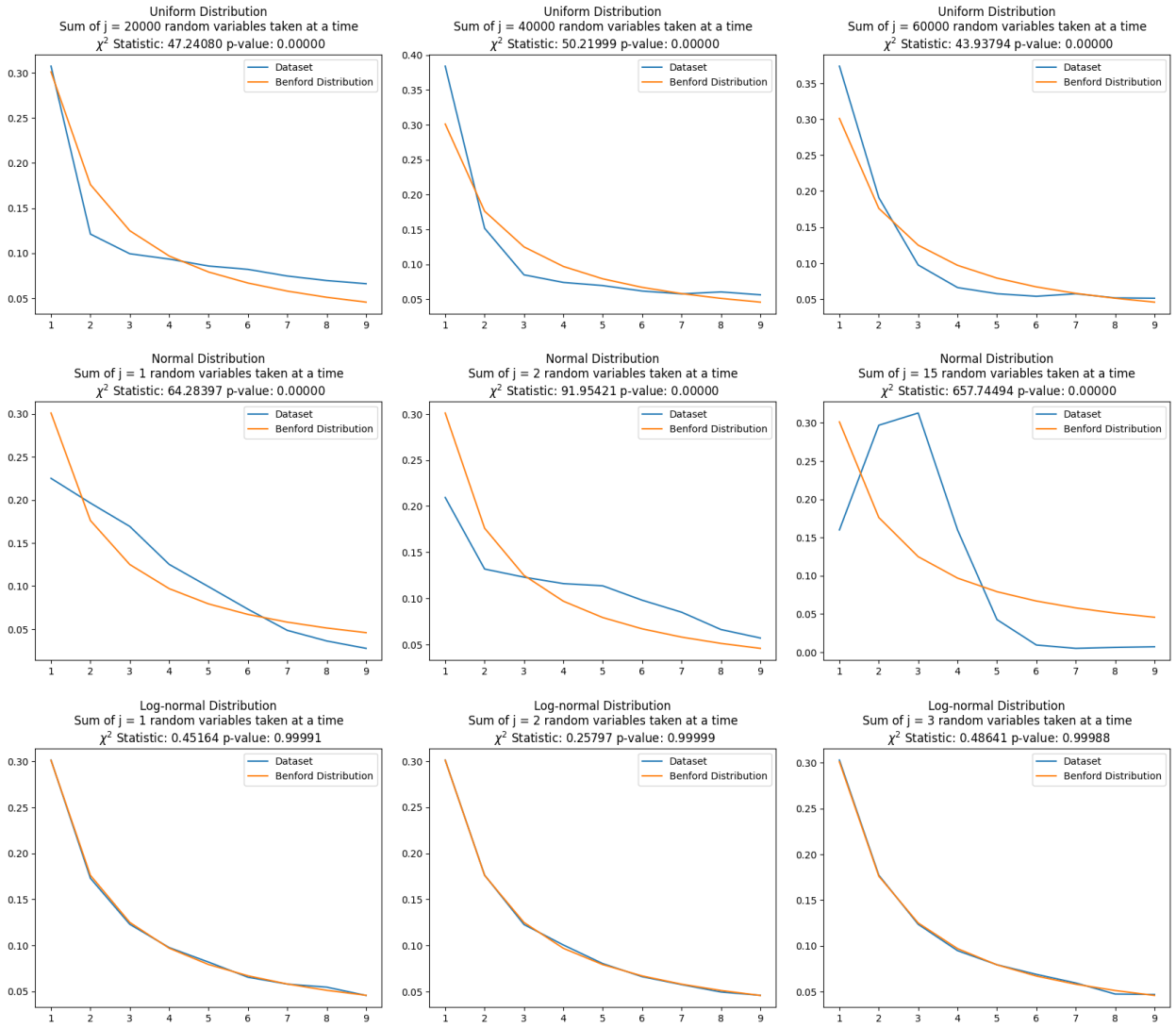

Figure 2: Taking the sum of j random variables at a time. The random variables are uniformly distributed in the 1st row, normally(Gaussian) distributed in the 2nd row and log normally distributed in the 3rd row.

3.3 Sine

In figure 3, we see the plot of the frequency of 1st digit of the $\sin^k(x)$ of the random variables repeated j times. Here we have taken k ranging from 1 to 4. We have generated the random variables using the uniform distribution. In the first case we see that it does not tend to follow Benford's law as j increases. We see that the p values are < 0.05 which implies that the null hypothesis (data follows Benford's law) can be rejected and the alternate hypothesis (data does not follow Benford's law) is supported by the data. In the second to fourth case we see that it tends to follow the Benford's law as j increases and that $p > 0.05$. We should note that these values become very small quickly and start causing floating point errors so we cannot perform too many iterations.

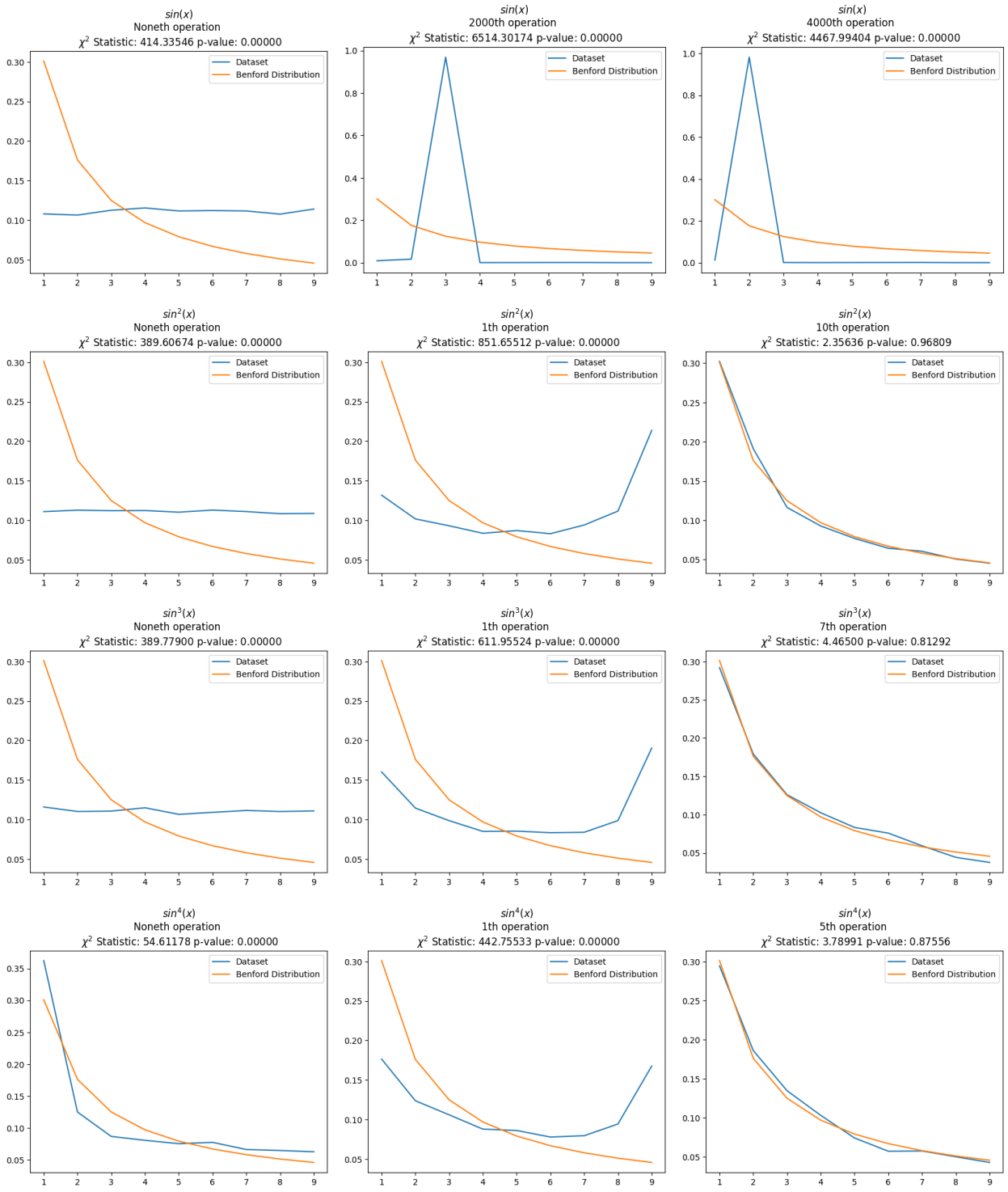

Figure 3: Performing $\sin^k(x)$ j times on uniformly distributed random variables. Here k varies depending on the row number ($k = 1$ for the 1st row, $k = 2$ for the 2nd row, $k = 3$ for the 3rd row and $k = 4$ for the 4th row.

3.4 Logarithm

In figure 4, we see the plot of the frequency of 1st digit of the log of the random variables repeated j times. We have generated the random variables using the uniform distribution. We see that it does not tend to follow Benford's law as j increases. We see that the p values are < 0.05 which implies that the null hypothesis (data follows Benford's law) can be rejected and the alternate hypothesis (data does not follow Benford's law) is supported by the data. We should note that these values become very small quickly and start causing floating point errors so we cannot perform too many iterations.

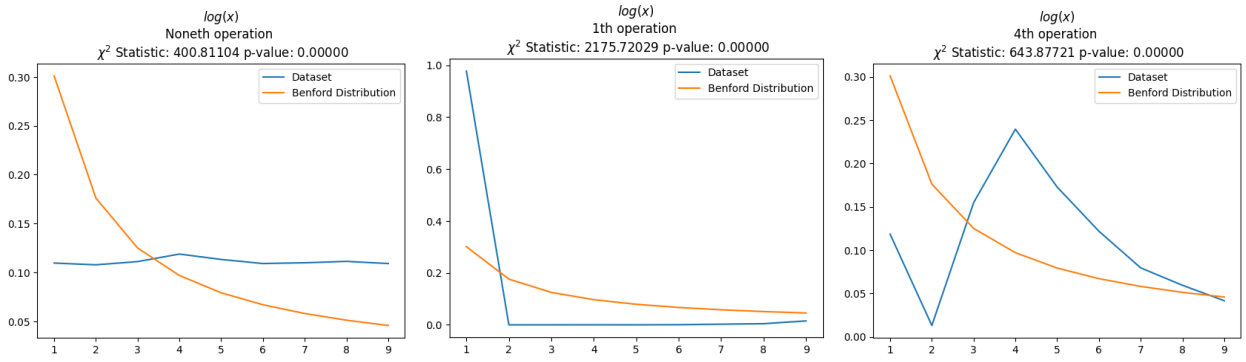

Figure 4: Performing $\log(x)$ j times on uniformly distributed random variables

3.5 Exponent

In figure 5, we see the plot of the frequency of 1st digit of the e^x of the random variables repeated j times. We have generated the random variables using the uniform distribution. We see that it does not tend to follow Benford's law as j increases. We see that the p values are < 0.05 which implies that the null hypothesis (data follows Benford's law) can be rejected and the alternate hypothesis (data does not follow Benford's law) is supported by the data. At $j = 1$ we should note that it fits pretty well with Benford's law. This is expected since exponential random variables tend to follow Benford's law [15]. We should also note that these values become very large quickly and start causing overflow errors so we cannot perform too many iterations.

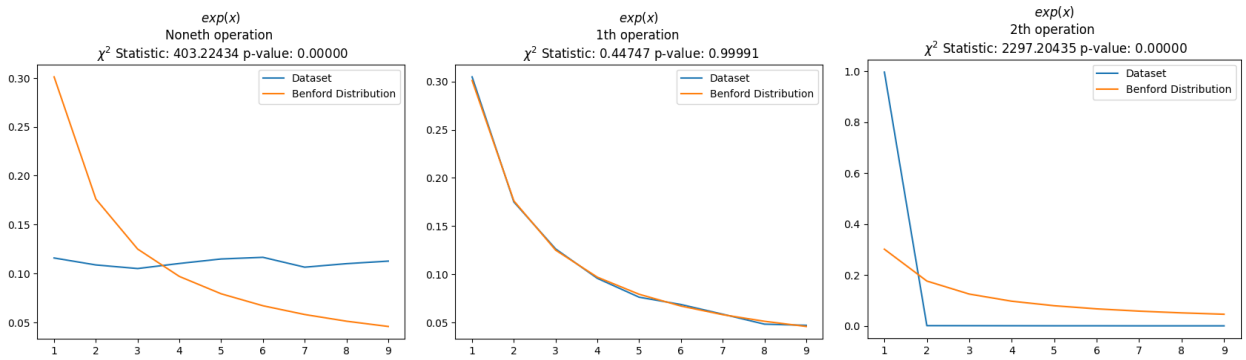

Figure 5: Performing e^x j times on uniformly distributed random variables

4 Conclusion

In this project we have simulated the products of random variables and have seen that they tend to follow Benford's law as we increase the number of random variables taken at a time. This is as expected since the $\log_B |X_1 \dots X_M| \mod 1$ tends to follow the uniform distribution [1]. We have also explored the results of how other arithmetic operations like sums, sine, logarithm and exponents. The code for this project can be found at [14].

5 Acknowledgements

We would like to thank Prof. Shantanu Desai for giving us this wonderful opportunity to credit the project work we have done and for teaching the concepts in class.

6 New things learnt

1. Learnt about Benford's law and the χ^2 test.
2. Learnt about implementing code to check for probabilities and running simulations.
3. Read literature on statistical methods
4. Learnt how to collaborate using Git-GitHub.
5. Improved coding with python and learnt to apply various python libraries.

REFERENCES

[1] Steven J Miller and Mark J Nigrini. The modulo 1 central limit theorem and benford's law for products. *arXiv preprint math/0607686*, 2006.

[2] Arno Berger and Theodore P Hill. Benford's law strikes back: No simple explanation in sight for mathematical gem. *The Mathematical Intelligencer*, 33(1):85, 2011.

[3] H Sakamoto. On the distributions of the product and the quotient of the independent and uniformly distributed random variables. *Tohoku Mathematical Journal, First Series*, 49: 243–260, 1943.

[4] Melvin Dale Springer and WE Thompson. The distribution of products of independent random variables. *SIAM Journal on Applied Mathematics*, 14(3):511–526, 1966.

[5] AK Adhikari. Some results on the distribution of the most significant digit. *Sankhyā: The Indian Journal of Statistics, Series B*, pages 413–420, 1969.

[6] AK Adhikari and BP Sarkar. Distribution of most significant digit in certain functions whose arguments are random variables. *Sankhyā: The Indian Journal of Statistics, Series B*, pages 47–58, 1968.

[7] Persi Diaconis. The distribution of leading digits and uniform distribution mod 1. *The Annals of Probability*, 5(1):72–81, 1977.

[8] Karl Pearson. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, 185:71–110, 1894.

[9] Karl Pearson. X. contributions to the mathematical theory of evolution.—ii. skew variation in homogeneous material. *Philosophical Transactions of the Royal Society of London.(A.)*, (186): 343–414, 1895.

[10] Karl Pearson. Xi. mathematical contributions to the theory of evolution.—x. supplement to a memoir on skew variation. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 197(287-299):443–459, 1901.

[11] Karl Pearson. Ix. mathematical contributions to the theory of evolution.—xix. second supplement to a memoir on skew variation. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 216(538-548): 429–457, 1916.

[12] William G Cochran. The $\chi 2$ test of goodness of fit. *The Annals of mathematical statistics*, pages 315–345, 1952.

[13] Ronald A Fisher. On the interpretation of χ 2 from contingency tables, and the calculation of p. *Journal of the royal statistical society*, 85(1):87–94, 1922.

[14] Vibhavasu Pasumarti and Raghav Juyal. Code for Benford Law Project. `https://github.com/DarkWake9/Project-Benford`, 2023.

[15] Anton K Formann. The newcomb-benford law in its relation to some common distributions. *PloS one*, 5(5):e10541, 2010.