# DATA MINING FOR BENFORD'S LAW IN ANCIENT ROMAN COINS

Akhilesh Warty and Eugene Pinsky

Metropolitan College, Boston University, 1010 Commonwealth Avenue, Boston, MA 02215, USA

## *ABSTRACT*

*This study examines the application of data mining techniques to analyze ancient Roman coin datasets and investigate the extent to which Benford's Law is exhibited in the numerical values of the coins. Ben- ford's Law predicts the frequency distribution of the first digits in naturally occurring datasets, and its applicability has been demonstrated across diverse fields. This research aims to explore whether ancient Roman coin values conform to this mathematical phenomenon, providing insights into the au- thenticity and naturalness of the data. By employing data mining methods, we analyze the leading digit distribution in a comprehensive dataset of ancient Roman coins. Additionally, we investigate trends and features within the dataset, such as coin composition, weight and diameter. The findings of this study contribute to the broader understanding of historical numismatic data and the relevance of Benford's Law in historical datasets.*

## *KEYWORDS*

*Numismatics, Coin Classification, Benford's Law, Data Mining*

## 1. INTRODUCTION

In this study, we scraped a comprehensive data set [13] that contains coins from the Roman Republic, the Roman Empire, and the Kingdom of Macedonia, capturing key characteristics such as denomination, weight, mint, ruler, and material. The collected data was then cleaned and structured for analysis using standard analysis techniques [10]. Preliminary exploratory data analysis (EDA) was performed to gain initial insights [6] into the data set before investigating the Benford's Law in the dataset. This included examining the distribution of different coin types, identifying trends in coin characteristics [1]. Through this analysis, our objective was to uncover patterns and anomalies that may guide further research into the numismatic history represented by the dataset.

Historically, other datasets have been examined and have shown to exhibit Benford's Law. The datasets range from atomic weights all the way to analyzing natural disasters. Benford's Law is also used as a technique when it comes to fraud detection. In a scientific report [8] written in 2015, Benford's law was used to examine and investigate the homogeneity of natural hazards to improve the detection of cyclones in the future. Additionally in 2011, Information Systems Audit and Control Association (ISACA) published an article [7] that explained the importance of Benford's Law and its use as a potential technique in various datasets to check for its authenticity. In 1999 an article was published [11] that described the mathematical phenomenon of Benford's Law and how it can be used to uncover irregularities and be used by CPA's. The same author also published a book [12] a book about the importance of Benford's Law as forensic accounting tool to analyze and authenticate data.

DOI:10.5121/ijdkp.2025.15203

This paper will investigate if the natural mathematical phenomenon is present in the ancient roman coins in its various features.

## 2. WHAT IS BENFORD'S LAW?

Benford's Law or Bernoulli's Law describes a fascinating phenomenon in the distribution of leading digits in numerical datasets. According to this law, in many real-world datasets, the first digit is more likely to be a smaller number, with 1 appearing as the leading digit about 30% of the time, and each subsequent digit becoming progressively less common [19]. It is defined using a mathematical theorem defined as:

$$P(d) = log_{10}\left(1 + \frac{1}{d}\right), \qquad d \in \{1, \ldots, 9\} \qquad (1)$$

This counterintuitive distribution holds for a wide range of datasets, including financial records, population numbers, and physical constants. This paper will examine and investigate if the ancient coins and monetary system during the Ancient Roman Empire, Roman Republic and Macedonian Empire (Which was part of the Roman Empire) seems to exhibit this natural law.

## 3. INITIAL EXPLORATORY DATA ANALYSIS

Before conducting the Exploratory Data Analysis (EDA), several key steps were carried out to prepare the dataset. These steps are outlined below:

- **Data Mining:** The data is mined [4] [5] using effective techniques [14] to create a comprehensive dataset.
- **Data Conversion:** The mined data is converted into a homogeneous structure, typically a Comma Separated Value (CSV) file, to facilitate further analysis.
- **Data Ingestion:** The CSV file is ingested into the analysis environment.
- **Data Transformation:** The data undergoes various transformations to ensure consistency and us- ability for analysis.
- **Creating a Subset:** The dataset is filtered [16] to include relevant records, such as those from the Ancient Roman Empire, Roman Republic, and Macedonian Empire.
- **Exploratory Data Analysis (EDA):** Initial analysis is conducted to gain an understanding of the dataset and identify underlying relationships [9].

The initial step of the entire process was data mining. This was done to extract public information available [13] into local storage for investigation and analysis.
After the data mining the information was then stored in a CSV file format for easier usage with the common data analysis packages such as Pandas or NumPy. After which the data was ingested into the analysis environment for examination and analysis.

The dataset that was mined is extremely extensive and included multiple countries, kingdoms, empires throughout history. Like every other dataset, the data is prone to missing values and outliers. Since the data is mined, it is prone to errors in the string and numeric conversion. The values were converted from a mixture from strings and Unicode to numeric data. To deal with the missing records while protecting its integrity, the missing observations were dropped. Common strategies include filling the data with the mean or median of the data depending on the attribute's distribution. Alternatively, the records can be dropped as well. This can be appropriate in certain scenarios.

The rationale behind dropping the data is due to the unique nature of the ancient coins:

- Unique Nature of Each Coin
- Historical Accuracy Of the dataset
- Preserving Data Integrity

The coins are unique and have immense historical significance, and the features in the dataset reflect this nature. Filling in the values by mean or median can cause occurrences that might not be possible, especially since the data is examined over an extensive period. To preserve the integrity, uniqueness, and historical nature of the data, missing records were deleted. Although this approach can commonly lead to sparse datasets, even after this data cleaning process, there were more than 32,000 coins present in the dataset without missing features.

To get an initial understanding and examine the underlying relationships present in the dataset, Exploratory Data Analysis is conducted [9]. After cleaning the data, finding missing records, the dataset is filtered [16] to include the Ancient Roman Empire, Roman Republic and Macedonian Empire (Which was part of the Roman Empire). Some initial distributions and observations are made to make informed approaches in the data analysis uses to extract information about its possible adherence to Benford's Law.

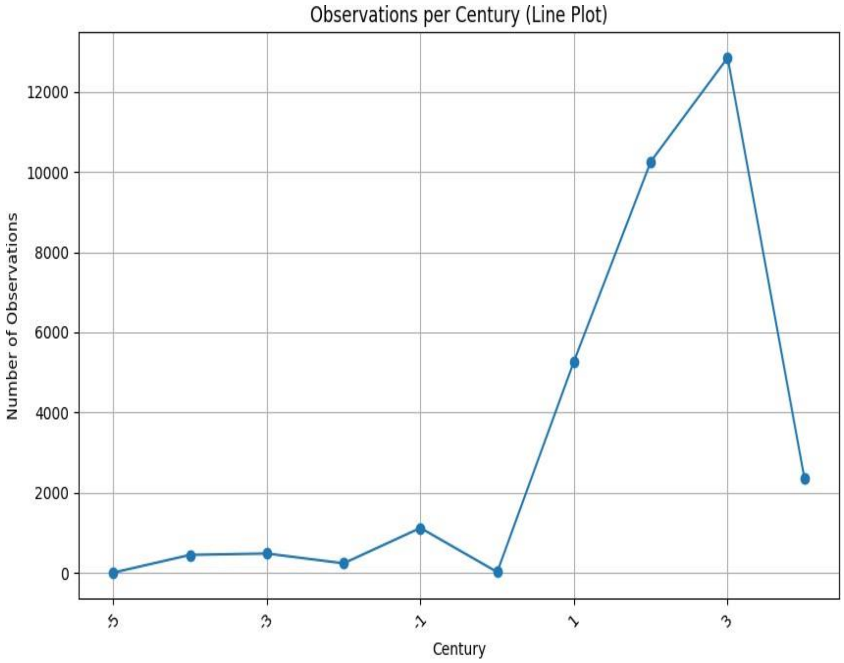


Figure 1: Coins Per Century

As seen in Figure 1, the observations are noted, and a period is calculated in the form of Centuries. The examined period spans from 5th Century BC till 4th Century AD. This period encompasses the Ancient Roman Empire as well as the Roman Republic. This is key since it gives a numerical base line to group and categorize the coins per century to examine each one of them on their own.

Since the goal is to examine if the monetary system exhibited Benford’s Law, we choose to examine not only the denominations [18] which will be expanded in the section below but also other characteristics as metal composition, coin diameter, coin weights etc. This allows us to view not only the monetary changes in the system but also to view the coins as natural entities. Since Benford’s law is said to be exhibited in many wide arrays of datasets, examining the multifaceted nature of the coin gives a holistic approach.

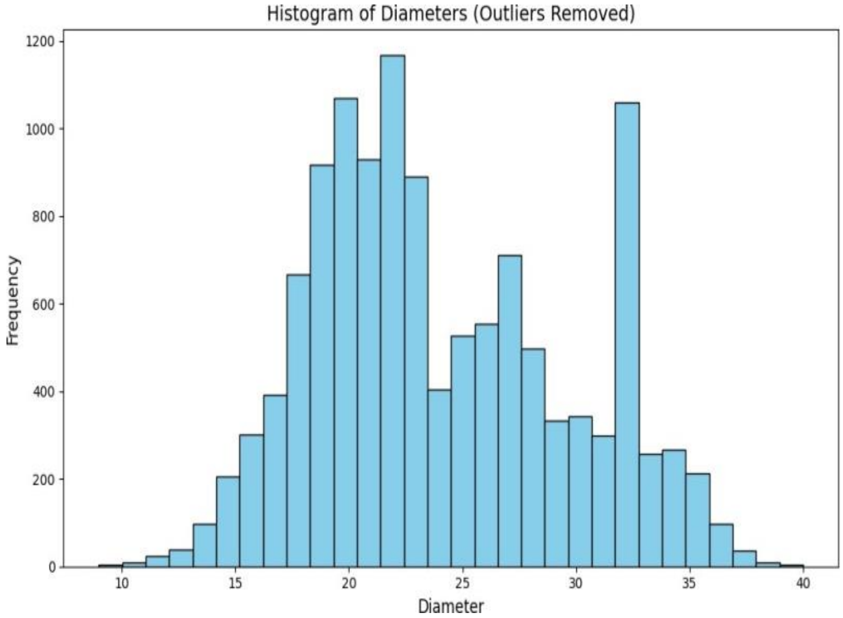


Figure 2: Distribution Of Coin Diameters

Figure 2 displays the frequency distribution of diameters, with outliers removed [10] [15]. The distribution appears roughly bimodal, with peaks around 20 and 32 units. The majority of diameters fall within the range of 17 to 35 units. Although there is a noticeable difference in frequency between the two peaks, the distribution remains continuous. The data exhibit some skewness to the right, indicating a slightly longer tail on the higher end of the diameter values [2].

Outliers were removed from the dataset by examining the distribution and looking for any discontinuous data points. This is done to remove the outliers of the data to get a better understanding of the distribution of the diameters in the dataset.

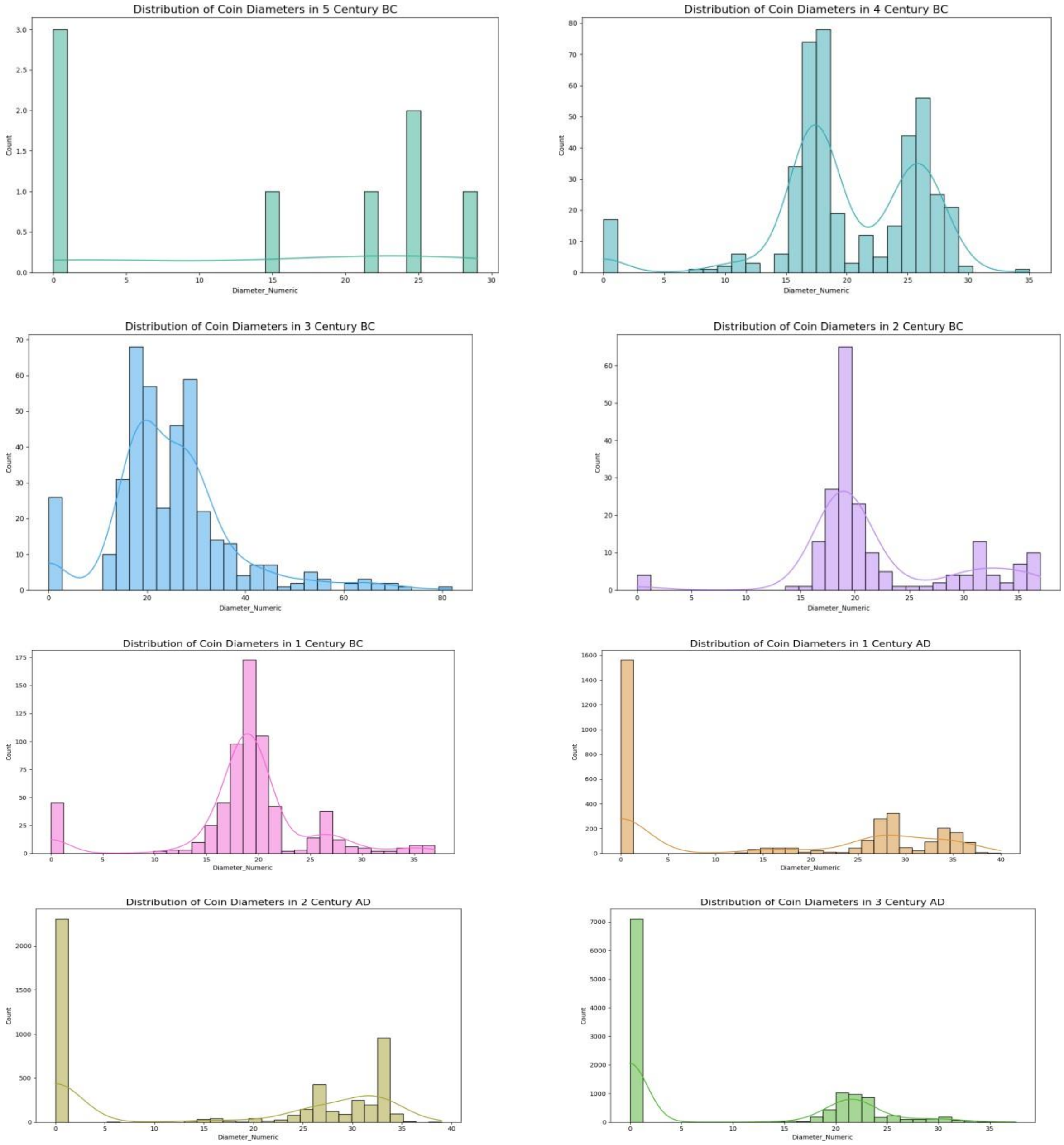


Figure 3: Diameter Distribution of Ancient Coins: 5th Century BC to 4th Century AD

Figure 3 examines the trends of the coin diameters that are parsed from strings literals to integers. Over time, the coin in a certain characteristics and distributions that share some similarities as well as some stark differences

In the 5th Century BC, the distribution of coin diameters was quite sparse, resembling a discrete distribution than a continuous one. There are peaks round 15 mm, 22 mm, and 25 mm and 28 mm. This can be attributed to the sparse information about the coins in the dataset or less variety in diameters during this period. Coin production was likely limited, reflecting a less developed economy or fewer resources dedicated to minting.

The 4th Century BC exhibits a continuous distribution which is a stark difference from the previous century, with a noticeable peak around 18 mm. This suggests an increase in the variety and number of coins produced compared to the previous century. The economy might have been growing, leading to a greater need for diverse coinage.

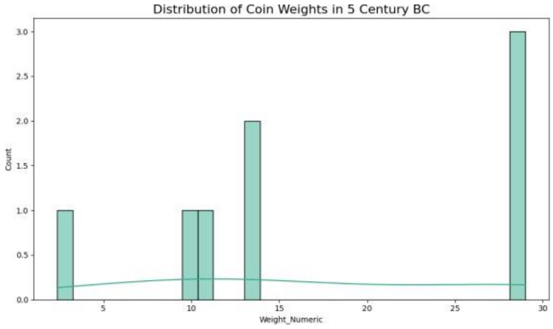
During the 3rd Century BC, the distribution has shifted to the right. There is a significant peak around 20 mm, with a broader spread of diameters, indicative of a higher number of coins, reflecting the growth of the economy. During this period was the Punic wars [3] which led to the Roman economy becoming stronger due to the capture of more territories, increase in the trade routes and a supply of more raw materials and metals which were used for the coins.

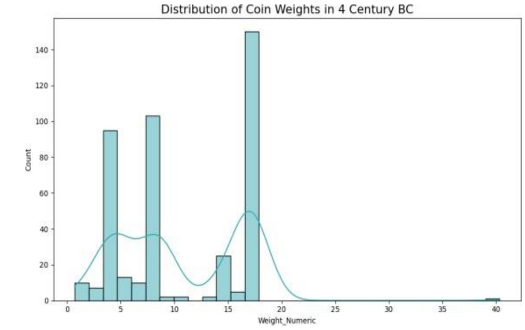
The 2nd Century BC featured a sharp peak of around 20 mm, with fewer coins in other diameter ranges. This more concentrated distribution that resembles the distribution of the coins during the 4th Century BC before the Punic wars [3]. During this period, the 3rd Punic wars [3] was ongoing which could play a part in changing the raw materials to fight the war.

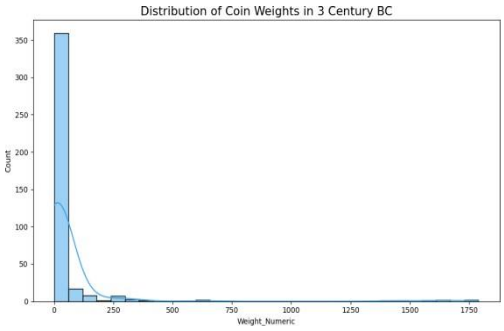
In the 1st Century BC, there was a prominent peak of around 20 mm, with a secondary peak around 26 mm. The higher numerical count could allude to the fact that the more territories were captured during the 3rd Punic war [3] leading to a more econometric shift for Rome.

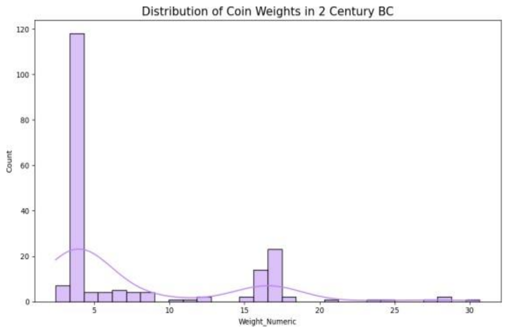
The 1st Century AD showed a significant peak around 20 mm, with a broader spread of diameters. The noticeable increase in the count of coins indicates a high production rate, reflecting economic growth and stability. The variety in coin sizes suggests an adaptable minting process to meet various demands. There is a big shift in the counts of the coins minted during this time period since the number of coins are at an all-time high.

During the first three centuries in the Anno Domini, the distributions are like each other, during the 1st century AD has a little more spread than the other two centuries. During the second century the coins are showing trends of becoming closer together in the diameter size. In the 3rd Century, Diocletian led the roots for standardization of the monetary system that could explain the increase in the counts as well as the less spread in the coin size present [2] [6].









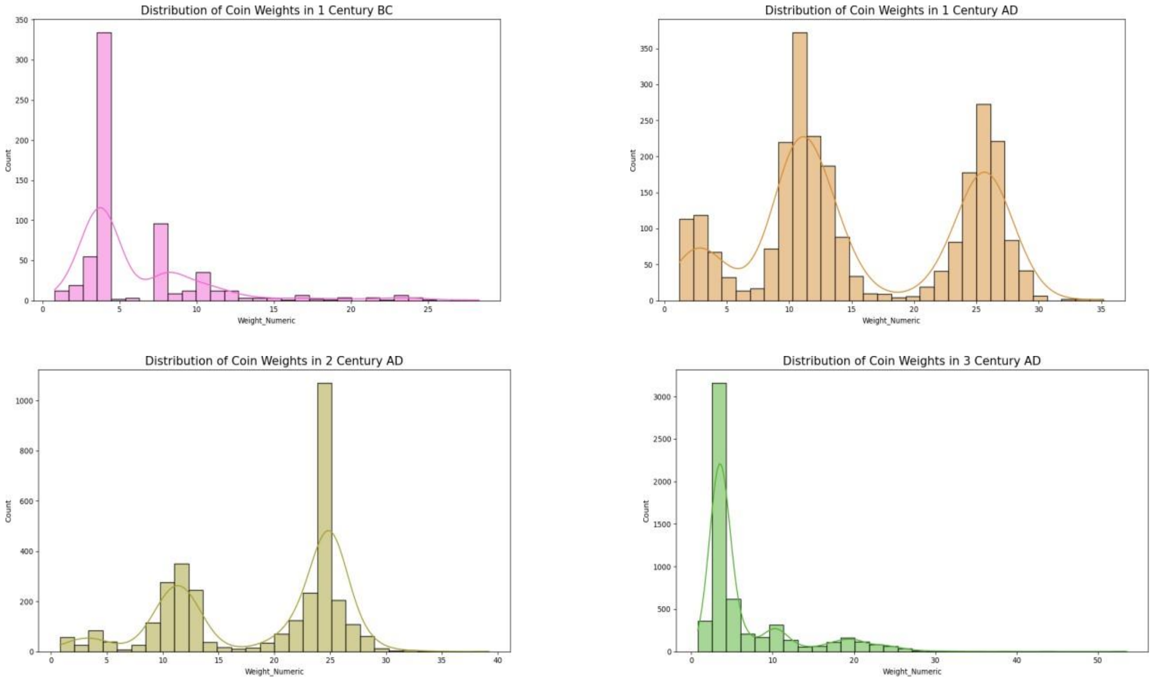


Figure 4: Weight Distribution in Ancient Roman Coins: 5th Century BC to 4th Century AD

Figure 4 examines the trends of the coin weights that are parsed.

In the 5th Century BC, the distribution of coin diameters was quite sparse, resembling a discrete distribution than a continuous one. There are peaks round 2 g, 10 g, and 13 g and 28 g. This is due to the sparse information that was mined during that era.

The 4th Century BC saw a more continuous distribution of coin diameters, with a noticeable mean around 10 to 15 g. This is a change from the previous century and created trends that were then standardized and expanded upon as time passed.

During the 3rd Century BC, there was a significant peak around 4 g, with a broader spread of diameters. The anomaly is the presence of certain coins that are usually larger than the normal set before in the previous periods.

The 2nd Century BC featured a sharp peak of around 4 g, with fewer coins in other diameter ranges. This is the period where the distribution that laid a more concrete foundation that were built upon in the next centuries and the progression of the Punic [3] war.

In the 1st Century BC, there was a prominent peak of around 4 g. The distribution is shifted to the right since the peaks are on the smaller size and there are very less counts and outliers compared to the previous century. A larger count suggests a integration of more coins and resources that were obtained during the century.

The 1st Century AD coins display a very different trend and shows multi modal peaks and a more even continuous distribution [2]. The counts of these occurrences are very high and shows that the there are three sets of coins that are being used. Which could have prompted the use of standardization in the next century by Diocletian. The next centuries show changes to this distribution and in the 2nd century AD still shows some distinct groups of coins which is then refined into a more consolidated distribution with all the occurrences becoming standardized around one group of coinage/unit.

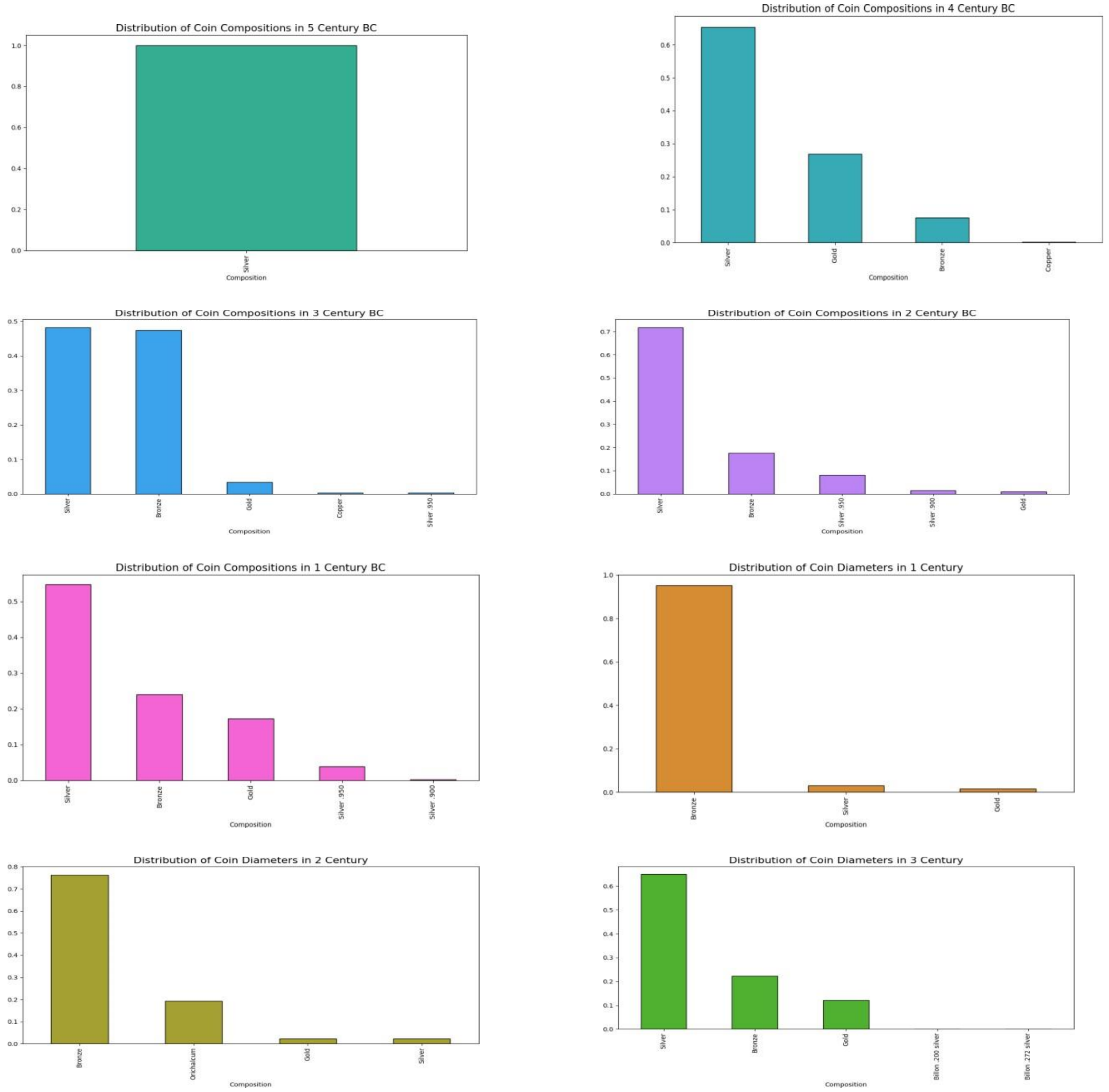


Figure 5: Composition Distribution in Ancient Roman Coins: 5th Century BC to 4th Century AD

In the 5th Century BC, the distribution of coin compositions was relatively sparse with silver being the only sort of composition used.

The 4th Century BC saw a more continuous distribution of coin compositions, with a noticeable increase in the variety of metals used. This most likely alludes to the growth of the economy as well as a need for a wider range of coins to keep up with the expansion of the economy and wealth. Silver is still by far the most used metal followed by gold. This graph does exhibit Benford’s law since the first occurrence is way higher than the second largest occurrence.

During the 3rd Century BC, the previous centuries efforts were added to since there is an increase in bronze production [17] and a reduction in gold coins. This was since silver is the most used currency till then and the gold coins were a rare occurrence and denoted extremely high amounts of wealth. This reduction of the occurrences of gold could be due to the data that might be available for that period.

The 2nd Century BC continues that trend however, there is a change in the volume of coins that were examined. This could be due to certain resources becoming more necessary during war to debase their currency and finance their military might.

In the 1st Century BC, there was a prominent peak in the distribution of coin compositions, with a higher count of coins. Silver became by far the most used metal for currency. The Denarius coin [18] became prevalent after roughly 211 BC and their circulation increased to the possible aftereffects of debasing their currency previously during the war.

The 1st Century AD shows a very different trend than the previous centuries. The widespread use of bronze instead of silver showed a change. The bronze coin was used for smaller denominations which were used for everyday purposes and the larger coins were used for trade. During this time the Roman Empire saw a significant increase in trade in its various territories as well as their expansion created after defeating Carthage during the Punic wars. The 2nd century AD still shows the same trend but has a new addition of which when examined had similar physical attributes to gold as well as the affordability to produce made it a popular metal for currency.

The 3rd Century AD shows a change in the trends since Silver was used more than bronze and was the most prevalent metal than bronze for minting coins. This could be due to the expansion that Rome went through which afforded them new materials and precious metals which they could use to create their currencies. This was the creation of the Denarius coin.

In the 4th Century AD, bronze seems to have overtaken the silver coin due to its debasement that was carried out over the past years that started with emperor Nero and that made the bronze coins more important and abundant.

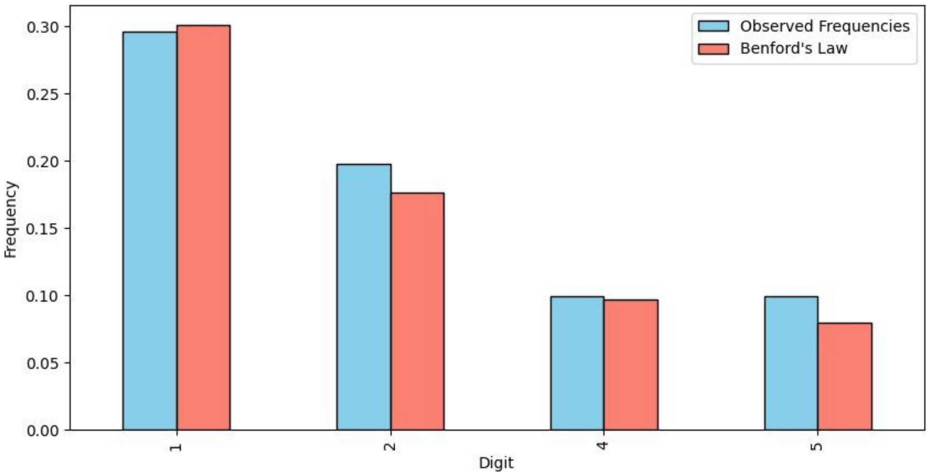


Figure 6: Benford’s Law for Ancient Coin Values

Figure 6 presents the frequency distribution for the numerical values for the entire dataset. This was done step by step so by first converting the numerous coinages into Denarius to establish a base line unit so that they can be compared against each other. After the conversion the values frequencies were normalized and noted so that they can be easily compared and examined to see if Benford’s law [19] is exhibited. By looking at the graph the phenomenon is seen since the first digit is more likely to be seen compared to the larger digits.

# 4. DATA

We start with the Frequency of last digits. This is shown in Table 1.

Table 1: Frequency of Last Digits Over Time Period

| **Last Digit** Start Century | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **8** |
|---|---|---|---|---|---|---|---|---|
| 5th Century BC | 11.11 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 100.00 | 0.00 |
| 4th Century BC | 18.52 | 0.00 | 33.33 | 100.00 | 0.00 | 2.08 | 0.00 | 100.00 |
| 3rd Century BC | 7.41 | 100.00 | 33.33 | 0.0 | 0.00 | 12.50 | 0.00 | 0.00 |
| 2nd Century BC | 14.81 | 0.00 | 33.33 | 0.00 | 100.00 | 8.33 | 0.00 | 0.00 |
| 1st Century BC | 7.41 | 0.00 | 0.00 | 0.00 | 0.00 | 16.67 | 0.00 | 0.00 |
| 1st Century AD | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 20.83 | 0.00 | 0.00 |
| 2nd Century AD | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 20.83 | 0.00 | 0.00 |
| 3rd Century AD | 11.11 | 0.00 | 0.00 | 0.00 | 0.00 | 14.58 | 0.00 | 0.00 |
| 4th Century AD | 29.63 | 0.00 | 0.00 | 0.00 | 0.00 | 4.17 | 0.00 | 0.00 |
| **Total** | **100.00** | **100.00** | **100.00** | **100.00** | **100.00** | **100.00** | **100.00** | **100.00** |

Table 1 looks at the frequency distribution of the last digits of some data across different centuries. The digit 0 is spread out over the century which is reflective of the monetary system and shows the roots of the contemporary monetary system. Whereas digits such as 5 are spread out over the centuries and are present consistently across the time period. This alludes to whether this set the roots for the numeric system that is used today where the values of 0,5 and 9 have grown from there.

Next, we show the frequency distribution of the last digits of some data across the time period in Table 2. Digit 0 is present across the period. 1 is also similarly spread out over the period. However, there are some stark differences that are observed such as 8,6,5 are present sparingly across this same time period. This is to be expected since the coins which had these numbers as their first digit to be present in their economy.

Table 2: Frequency of First Digits Over Time Period

| **First Digit** Start Century | **0** | **1** | **2** | **4** | **5** | **6** | **8** |
|---|---|---|---|---|---|---|---|
| 5th Century BC | 1.96 | 0.00 | 0.00 | 20.00 | 0.00 | 0.00 | 100.00 |
| 4th Century BC | 9.80 | 6.67 | 33.33 | 20.00 | 100.00 | 100.00 | 0.00 |
| 3rd Century BC | 15.69 | 6.67 | 33.33 | 20.00 | 0.00 | 0.00 | 0.00 |
| 2nd Century BC | 11.76 | 13.33 | 33.33 | 20.00 | 0.00 | 0.00 | 0.00 |
| 1st Century BC | 13.73 | 13.33 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 1st Century AD | 17.65 | 6.67 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 2nd Century AD | 15.69 | 13.33 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 3rd Century AD | 13.73 | 13.33 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 4th Century AD | 0.00 | 26.67 | 0.00 | 20.00 | 0.00 | 0.00 | 0.00 |
| **Total** | **100.00** | **100.00** | **100.00** | **100.00** | **100.00** | **100.00** | **100.00** |

## 5. EXHIBITING BENFORD'S LAW AND CONCLUDING REMARKS

Upon examining the composition of ancient Roman coins, it becomes evident that Benford's Law governs the numerical relationship between certain digits in naturally occurring datasets. Benford's Law applies to each digit in the decimal system, but the data shows frequencies only for some digits. Certain digits are not observed since the numerical values in the dataset do not contain them. For example, ancient Roman coinage values would not start with digits 3, 6, 8, or 9. However, the observed digits follow frequencies like those defined by Benford's Law. This is demonstrated in Figure 6.

In this paper, we investigated whether the characteristics of ancient Roman coins conform to Benford's Law. Specifically, we examined the first digit of the numerical values in the coinage. We found that these values partially exhibit Benford's Law because certain digits do not appear in the dataset.

Future research will extend the application of Benford's Law to other historical datasets.

## DECLARATIONS

**Conflict of Interest:** There are no conflicts of interest regarding the publication of this paper.

**Author Contributions:** All the authors contributed equally to the effort.

**Funding:** This research was conducted without any external funding. All aspects of the study, including design, data collection, analysis, and interpretation, were carried out using the resources available within the authors' institution.

**Data Availability (including Appendices):** All the relevant data, Python code for analysis, detailed annual tables and graphs are available via:
https://anonymous.4open.science/r/Numismatics-72C8/

### REFERENCES

[1] A. Agresti. *An Introduction to Categorical Data Analysis*. Wiley, 2007.
[2] V. Authors. *Statistics*. OpenStax, 2023.
[3] E. Britannica. Punic wars, 2025.
[4] J. Han, M. Kamber, and J. Pei. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers, 2012.
[5] D. Hand, H. Mannila, and P. Smyth. *Principles of Data Mining*. MIT Press, 2001.
[6] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2009.
[7] ISACA. Understanding and applying benford's law. *ISACA Journal*, 2011.
[8] R. Joannes-Boyau, T. Bodin, A. Scheffers, and et al. Using benford's law to investigate natural hazard dataset homogeneity. *Scientific Reports*, 5:12046, 2015.
[9] A. Maheshwari. *Data Analytics Made Accessible*. Springer, 2014.
[10] W. McKinney. *Python for Data Analysis*. O'Reilly Media, 2017.
[11] M. J. Nigrini. I've got your number. *Journal of Accountancy*, 187(5):79–83, May 1999.
[12] M. J. Nigrini. *Benford's Law: Applications for forensic accounting, auditing, and fraud detection*. John Wiley Sons, April 2012.
[13] Numista. Numista - the coin catalog, 2025.
[14] F. Provost and T. Fawcett. *Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking*. O'Reilly Media, 2013.
[15] C. R. Shalizi. *Advanced Data Analysis from an Elementary Point of View*. Unknown, 2012.
[16] S. Skiena. *The Data Science Design Manual*. Texts in Computer Science, 2010.

[17] L. University. Essay on the production of ancient coins, 2025.
[18] Vaulted. The roman denarius, 2025.
[19] Wikipedia. Benford’s law, 2025.