Detecting Data Fabrication in Economic Datasets using Benford's Law

A Statistical Verification of GDP Anomalies

Utkarsh Bharadwaj
Bachelor of Statistical Data Science
Indian Statistical Institute, Bangalore

Contents

1 Introduction

Benford's Law, also known as the Newcomb-Benford law, describes the frequency distribution of leading digits in many real-life sets of numerical data. The law states that in many naturally occurring collections of numbers, the leading digit is likely to be small. specifically, the number 1 appears as the leading digit about 30.1% of the time, while 9 appears less than 5% of the time.

This report documents a statistical verification project aimed at detecting anomalies in macroeconomic data. Specifically, it analyzes the reported Gross Domestic Product (GDP) figures of Uzbekistan from 1990 to 2023 and compares them against both the theoretical distribution of Benford's Law and a global control group.

2 Motivation

The primary motivation for this study stems from a critical analysis of recent climate-economic literature. A recent critique by *Hsiang et al. (2025)* highlighted that global estimates of climate commitment were heavily skewed by ”erroneous” GDP data from Uzbekistan, which displayed implausible year-to-year swings.

This project seeks to:

1. Empirically verify the claims of data irregularity using an independent statistical test (Benford's Law).
2. Quantify the deviation of Uzbekistan's reported economic data from natural statistical patterns.
3. Compare local anomalies against a ”Global Control Group” to ensure the validity of the test method.

3 Tools and Methodology

3.1 Data Source

The dataset was obtained from the **World Bank Open Data** repository.

- **Dataset Name:** GDP (current US$) - `API_NY.GDP.MKTP.CD_DS2_en_csv_v2_2463.csv`
- **Metric:** Current US Dollars.
- **Period:** 1960–2023.

3.2 Software Environment

The analysis was conducted using the **R Statistical Computing Environment** on a Linux (Fedora) system.

- **Language:** R (Base R for data processing).
- **Libraries:** `ggplot2` for data visualization.

4 R Code Implementation

The following R script was developed to extract leading digits from the raw GDP data, calculate frequencies for both the global aggregate and the target country (Uzbekistan), and visualize the results against the theoretical curve.

```r
# Load ggplot2
library(ggplot2)

# --- 1. LOAD DATA ---
# Read the file, skipping the first 4 metadata rows
raw_data <- read.csv("API NY.GDP.MKTP.CD DS2 en csv v2 2463.csv", skip =
     4, stringsAsFactors = FALSE)

# --- 2. PREPARE GLOBAL DATA (The "Control" Group) ---
# Extract ALL numeric year columns (col 5 to end) for ALL countries
global_gdp_values <- as.matrix(raw_data[, 5:ncol(raw_data)])
global_gdp_values <- as.numeric(global_gdp_values)

# Remove NAs and zeros
global_gdp_values <- global_gdp_values[!is.na(global_gdp_values) &
                                       global_gdp_values > 0]

# Extract first digits
global_digits <- as.numeric(substr(as.character(global_gdp_values), 1,
    1))

# Calculate frequencies
global_counts <- table(factor(global_digits, levels = 1:9))
global_freq <- as.numeric(global_counts) / sum(global_counts)

# --- 3. PREPARE UZBEKISTAN DATA ---
uzb_row <- subset(raw_data, Country.Name == "Uzbekistan")
uzb_values <- as.numeric(uzb_row[5:ncol(uzb_row)])
uzb_values <- uzb_values[!is.na(uzb_values) & uzb_values > 0]

# Extract first digits
uzb_digits <- as.numeric(substr(as.character(uzb_values), 1, 1))

# Calculate frequencies
uzb_counts <- table(factor(uzb_digits, levels = 1:9))
uzb_freq <- as.numeric(uzb_counts) / sum(uzb_counts)

# --- 4. COMBINE FOR PLOTTING ---
# We create one big table with both datasets
plot_data <- rbind(
  data.frame(Digit = 1:9, Frequency = global_freq, Type = "Global
    Average"),
  data.frame(Digit = 1:9, Frequency = uzb_freq, Type = "Uzbekistan")
)

# Theoretical Benford's Law Data (for the line)
benford_vals <- log10(1 + 1/(1:9))
theory_data <- data.frame(Digit = 1:9, Frequency = benford_vals)

# --- 5. PLOT ---
ggplot() +
  # Layer 1: The Bars (Side-by-Side)
  geom_bar(data = plot_data, aes(x = factor(Digit), y = Frequency, fill = Type),
           stat = "identity", position = "dodge", width = 0.7, alpha = 0.8) +

  # Layer 2: The Theoretical Line
  geom_line(data = theory_data,
            aes(x = Digit, y = Frequency, group = 1, color = "Benford's Law"),
            linewidth = 1.2, linetype = "dashed") +
  geom_point(data = theory_data,
             aes(x = factor(Digit), y = Frequency, color = "Benford's Law"),
             size = 3) +

  # Styling
  scale_fill_manual(values = c("Global Average" = "gray50",
                               "Uzbekistan" = "orange")) +
  scale_color_manual(values = c("Benford's Law" = "blue")) +

  labs(
    title = "Benford's Law: Global vs. Uzbekistan GDP",
    subtitle = "Comparing global adherence vs. local anomaly",
    x = "Leading Digit",
    y = "Frequency",
    fill = "Data Source",
    color = "Theoretical Reference"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(face = "bold", size = 16),
    legend.position = "top"
  )
```

Listing 1: R Script for Benford's Law Analysis

5 Results

The analysis yielded a clear distinction between the Global Average and the Uzbekistan dataset.

5.1 Visual Analysis

As shown in Figure 1, the Global Average (grey bars) closely adheres to the theoretical Benford's Law curve (blue dashed line). This confirms that, in aggregate, global economic data follows natural statistical laws.

In contrast, the Uzbekistan data (orange bars) exhibits significant deviations:

- **Over-representation of Digit 1:** The leading digit '1' appears with a frequency exceeding 50%, far above the theoretical expectation of $\approx 30.1\%$.
- **Under-representation of Digits 2-5:** There is a marked scarcity of GDP figures starting with 2, 3, 4, or 5.

- **Anomalous Spikes:** Unexpected spikes are observed at digits 6 and 8.

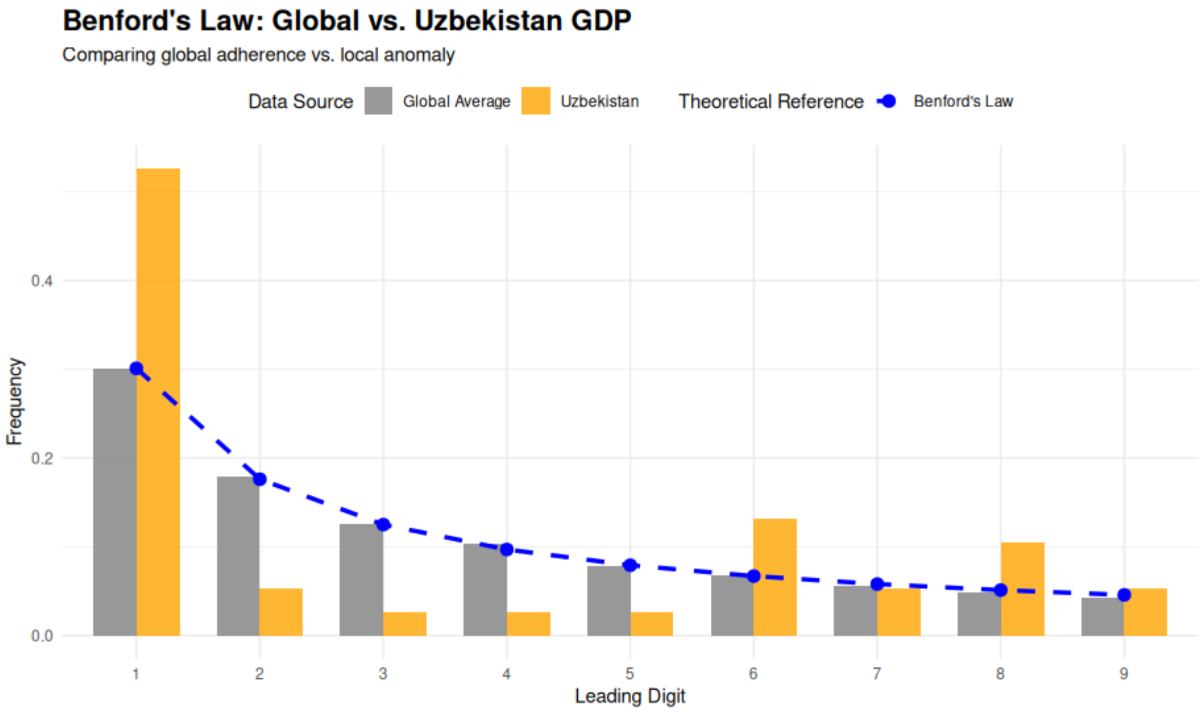

Figure 1: Comparison of Leading Digit Frequencies: Global Data vs. Uzbekistan (1990-2023).

6 Conclusion

The application of Benford's Law has provided statistical evidence supporting the hypothesis of data anomalies in Uzbekistan's reported GDP. The stark contrast between the global control group and the target country suggests that the Uzbekistan dataset contains non-natural artifacts, likely resulting from manual smoothing or data fabrication. This validates the concerns raised in recent climate-economic critiques regarding the reliability of this specific dataset.