Presented by Team - Survey Corps

BENFORD'S LAW ANALYSIS ON TWITTER DATA

Maths & AI Project

Team Members

Shriti Negi
Sayuri Janbandhu
Devaansh Kathuria
Geethika

Code Contributors

Code

Shriti, Geethika – Led the development of core notebook code and implementation.

Analysis

Sayuri– Provided insights, interpretations, and data-driven conclusions.(Also played a major role in designing and creating this presentation.)

Debug

Devaansh – Identified and resolved critical bugs to ensure smooth execution.

Presentation Contributors

Design

Shriti – Created the overall visual theme, layout, and slide aesthetics.

Editing

Devaansh, Geethika – Refined the content, ensured clarity, and maintained consistency throughout the presentation.

Content

Sayuri - Wrote and structured the main content, aligning it with the project's goals.

Introduction to Benford's Law

🔍 **What is Benford's Law?**

Benford's Law predicts that in many naturally occurring datasets, the leading digit is more likely to be small.

Distribution:

1 appears ~30% of the time

9 appears <5% of the time

Used in forensic accounting, fraud detection, and anomaly detection.

"Benford's Law helps us detect whether a dataset behaves 'naturally' or shows signs of manipulation."

Project Objective & Dataset Overview

Aim:

- *To verify whether Benford's Law holds for numerical data in a real-world Twitter dataset.*

Dataset:

- *CSV file sourced from Twitter data.*
- *Key numerical columns analysed:*
  - *id*
  - *followersCount*
  - *friendsCount*

Key Questions:

- *Do these columns follow the expected digit distribution according to Benford's Law?*
- *Which features show the most or least conformity?*

Why Twitter Data?

- *Massive volume and naturally generated user metrics.*
- *Diverse behavior patterns make it ideal to test Benford's hypothesis.*

How We Tested Benford's Law

```python
import numpy as np
from scipy.stats import chisquare

# Extract leading digit
digits = df['followersCount'].dropna().astype(str)
digits = digits.str.lstrip('-').str.replace('.', '').str[0].astype(int)

# Count observed frequencies
observed = [digits.tolist().count(d) for d in range(1, 10)]

# Expected frequencies (Benford's Law)
expected = [np.log10(1 + 1/d) * len(digits) for d in range(1, 10)]

# Chi-square test
chi2, p_value = chisquare(observed, f_exp=expected)
```

1. Extract the first digit of each numeric value
2. Count how often each digit (1–9) appears
3. Calculate expected frequencies using Benford's formula:
    - $P(d) = \log_{10}(1 + 1/d)$
4. Run a Chi-square goodness-of-fit test
5. Plot observed vs expected frequencies

Data → Digit Extraction → Frequency → Chi^2 Test → Plot

📉 Observed vs Expected Results

Column	Chi² Value	p-value	Fit Quality
id	27.71	0.0005	❌ Poor fit
followersCount	25.48	0.0013	⚠ Moderate fit
friendsCount	23.32	0.003	✅ Best fit

🟩 Which column matched Benford's Law best? **-friendsCount**

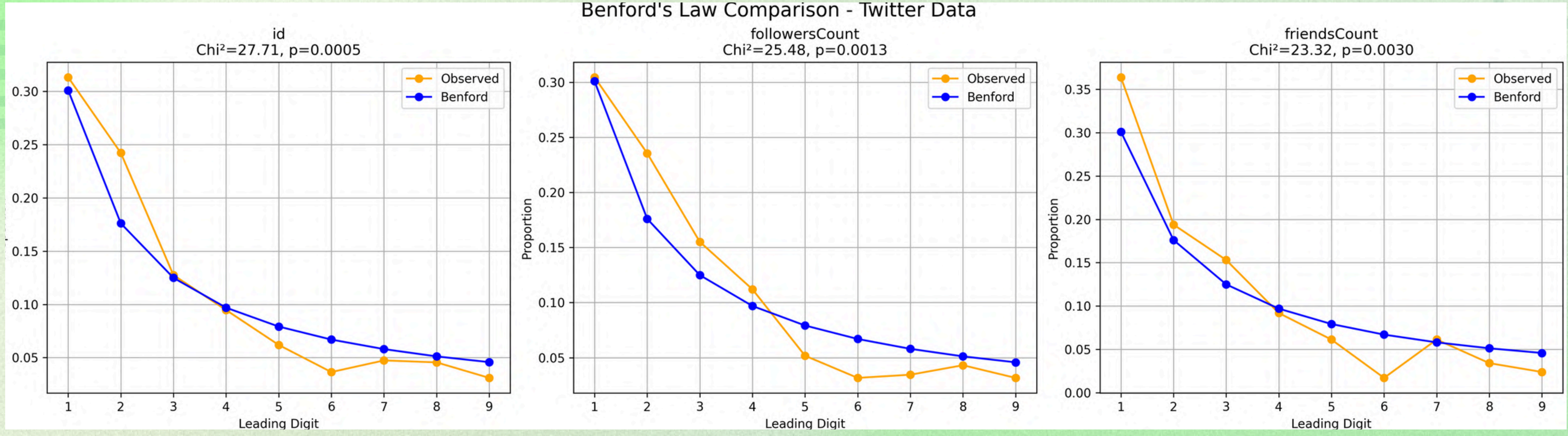
- The observed distribution (orange line) is quite close to the expected Benford curve (blue).
- Fewer sharp deviations in digit frequencies.
- Chi-square statistic = 23.32, p-value = 0.0030 (still significant, but lowest chi² among all).

🟥 Which deviated most? **-id**

- The orange line diverges the most from Benford's curve.
- Especially visible in digits 2–5.
- Chi-square statistic = 27.71, p-value = 0.0005
- Likely due to IDs being artificially generated or sequential, which violates Benford's assumptions.

🟨 Middle Ground **-followersCount**

- Has moderate deviation — better than id, worse than friendsCount.
- Chi-square = 25.48, p-value = 0.0013



What Does It Tell Us?

01.

Overall Conformity with Benford's Law

- All three numerical columns (id, followersCount, friendsCount) showed a general trend consistent with Benford's Law.
- However, none of the columns perfectly fit the expected distribution (all p-values < 0.05).

02.

Statistical Takeaway

- Chi-square test results for all columns were significant.
- Implies that although the shape resembles Benford's curve, true statistical conformity is not achieved.
- Lower Chi^2 value = better match → friendsCount is closest.

03.

Conclusion from Insights

- Columns like followersCount and friendsCount show partial natural compliance with Benford's Law — useful in fraud or anomaly detection.
- id does not represent organic data, confirming Benford's Law is best applied to naturally occurring datasets.

Conclusion & Future Scope

Conclusion

- Benford's Law provides a useful lens to analyze patterns in real-world numerical data.
- From the Twitter dataset:
  - friendsCount showed the best fit.
  - id diverged due to artificial generation.
- The results reinforce that Benford's Law works best with organically distributed, large-scale data.

🔭 ***Future Scope***

- 🔍 *Test additional Twitter features, e.g., retweet counts, likes, or tweet lengths.*
- 🧠 *Incorporate machine learning to detect anomalies or bots based on Benford deviation.*
- 🌍 *Apply Benford's Law to other social platforms or domains (finance, health, etc.).*
- 🧪 *Automate Benford-check pipelines for real-time fraud detection.*

🎓 ***Key Takeaway***

Benford's Law is a powerful statistical tool when used on the right kind of data — especially for detecting unnatural or manipulated patterns.

Benford's Law is a quiet reminder that even in the messiness of life, patterns often hide where we least expect them—in the very first digit.

THANK YOU

~Team ***SURVEY CORPS***