The TC Data 360 dataset and the Benford’s Law

FRANCESCO PORRO
UNIVERSITA’ DEGLI STUDI DI MILANO-BICOCCA
The TC Data 360 dataset - introduction

What is the TC Data 360 dataset?

“The TC data 360 is an initiative of the World Bank Group's Macroeconomics, Trade & Investment Global Practice, which helps countries achieve the Bank Group's twin goals, ending extreme poverty and boosting shared prosperity, through rapid and broad-based economic growth, centered on strong contributions from the private sector around the world”
What is the World Bank?

It is not a bank in the ordinary sense but a unique partnership to reduce poverty and support development. The World Bank Group comprises five institutions managed by their member countries.

Established in 1944, the World Bank Group is headquartered in Washington, D.C.

It has more than 10,000 employees in more than 120 offices worldwide.
Which kind of data are they?
The TC data 360 currently features over 2400 indicators for more than 200 countries in the time range 1960-2018:
Which kind of data are they?

- **Trade**
  - Trade Outcomes
  - Trade Barriers
  - Trade Facilitation
  - E-trade
  - ....

- **Investment**
  - Reform Progress (Investment)
  - Risk and Policy Uncertainty
  - Entry and Investment
  - ....

- **Innovation**
  - Innovation Outputs
  - Entrepreneurship
  - ....

- **Economy**
  - Economic Outcomes
  - Economic and Social Context
  - Monetary and Fiscal
  - ....

- **Sectors**
  - Manufacturing
  - Tourism (Sector)
The sources of the data

Where do the data come from?

The key sources of data on TCdata360 include:

- World Trade Organization
- World Economic Forum
- The World Bank
- International Monetary Fund
- International Trade Centre
- Global Entrepreneurship Monitor
- OECD

Porro, Francesco

The TC Data 360 dataset and the Benford’s Law
The main goal

Are there any variables following the Benford’s Law?

For same suitable variables all the available values in the dataset have been considered and a statistical test on the distribution of the first significant digit has been performed. The distribution target is:

**Benford’s Law:**

\[ P(D_1 = d_i) = \log(1 + \frac{1}{d_i}) \text{ where } d_i \in \{1, 2, ..., 9\} \]

**Frequency distribution:**

<table>
<thead>
<tr>
<th>values</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>frequencies</td>
<td>0.3010</td>
<td>0.1761</td>
<td>0.1249</td>
<td>0.0969</td>
<td>0.0792</td>
<td>0.0669</td>
<td>0.0580</td>
<td>0.0512</td>
<td>0.0458</td>
</tr>
</tbody>
</table>
The assessment of the fitting

It is well-known that the Pearson's $X^2$ test is consistent:

It is not useful for "big" samples

only variables with a limited number of observations

(<1600 obs)

have been considered
The assessment of the fitting

\[ X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \]

where:

- \( O_i \) is the observed frequency for the digit \( i \)
- \( E_i \) is the theoretical frequency of the digit \( i \), according to the Benford's Law

<table>
<thead>
<tr>
<th>( \alpha )</th>
<th>0.25</th>
<th>0.10</th>
<th>0.05</th>
<th>0.025</th>
<th>0.010</th>
<th>0.005</th>
<th>0.001</th>
</tr>
</thead>
<tbody>
<tr>
<td>Critical values: ( \chi^2_{8;1-\alpha} )</td>
<td>10.22</td>
<td>13.36</td>
<td>15.51</td>
<td>17.53</td>
<td>20.09</td>
<td>21.96</td>
<td>26.12</td>
</tr>
</tbody>
</table>

Porro, Francesco

The TC Data 360 dataset and the Benford's Law
Investment in water and sanitation with private participation (current US$)

Time range: 1987-2018

Total number of observations: $N=192$

Value of $X^2 = 2.786617$

P-value: $P(\chi^2_{8} > 2.786617) = 0.9470$

**Porro, Francesco**

The TC Data 360 dataset and the Benford’s Law
Investment in water and sanitation with private participation (current US$)

Time range: 1987-2018

Total number of observations: N=192

Value of $X^2 = 2.786617$

P-value $= P(X^2 > 2.786617) = 0.9470$

BENFORD’S LAW? YES

Porro, Francesco
The TC Data 360 dataset and the Benford’s Law
New businesses registered (number)

Time range: 2006-2015

Total number of observations: N=1293

Value of $X^2 = 7.980378$

P-value $= P(X^2 > 7.980378) = 0.4353$

BENFORD’S LAW?
New businesses registered (number)

Time range: 2006-2015

Total number of observations: \( N = 1293 \)

Value of \( X^2 = 7.980378 \)

P-value = \( P(X^2 > 7.980378) = 0.4353 \)

BENFORD’S LAW? YES
Investment in energy with private participation (current US$)

Time range: 2003-2010

Total number of observations: N=728

Value of $X^2 = 21.41651$

P-value $= P(X^2 > 21.41651) = 0.0613$

Benford’s Law?
Investment in energy with private participation (current US$)

Time range: 2003-2010

Total number of observations: N=728

Value of $X^2 = 21.41651$

P-value $= P(X_8^2 > 21.41651) = 0.0613$

**BENFORD’S LAW? NO**
Gross national expenditure (current US$)

Time range: 1961-2017

Total number of observations: \( N = 1564 \)

Value of \( X^2 = 256.3665 \)

P-value = \( P(X^2 > 256.3665) = 0 \)

**BENFORD’S LAW?**
Gross national expenditure (current US$)

Time range: 1961-2017

Total number of observations: N=1564

Value of $X^2 = 256.3665$

P-value = $P(X^2 > 256.3665) = 0$

BENFORD’S LAW? NO
The TC data 360 is a very huge dataset, provided by worthy and prestigious international institutions. There are more than 2400 indicators for more than 200 countries all over the world. The variables range in the time interval 1960-2018 (the availability varies). The variables are divided into five main areas: Trade, Investment, Innovation, Economy, and Sectors. The variables are very different: some of them are clearly not related to the Benford’s law, some are clearly related to it, and the all the other ones are suitable for an investigation... The number of the observations is an important issue to be taken under control!


