Do the data fit to Benford's law?

David Zelený

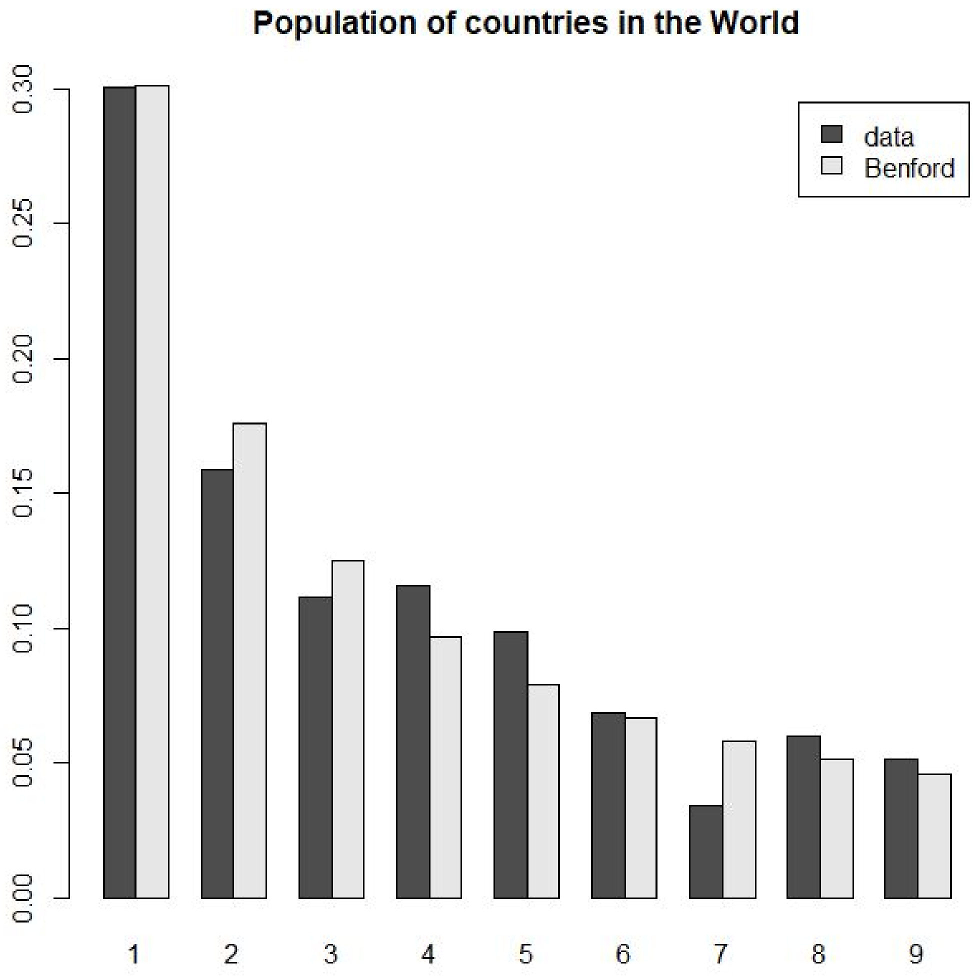
Introduction

Benford's law is an interesting mathematical phenomenon: in many real-world measured data sets, when we take the first digit of each measured value (i.e. the leftmost digit, regardless of the position of the decimal point), the most commonly occurring number will be 1, then 2 and so on, with decreasing frequency, with 9 being the least probable. See here for a nice website demonstrating Benford's law on different datasets, or Wikipedia for the explanation of why this is happening, or, e.g. this paper, where authors extracted correlation coefficients from many different scientific studies to check whether they follow Benford's law (and then asked researchers to 'fake' data, i.e. generate random correlation coefficients, to show that 'fake' data do not obey Benford's law).

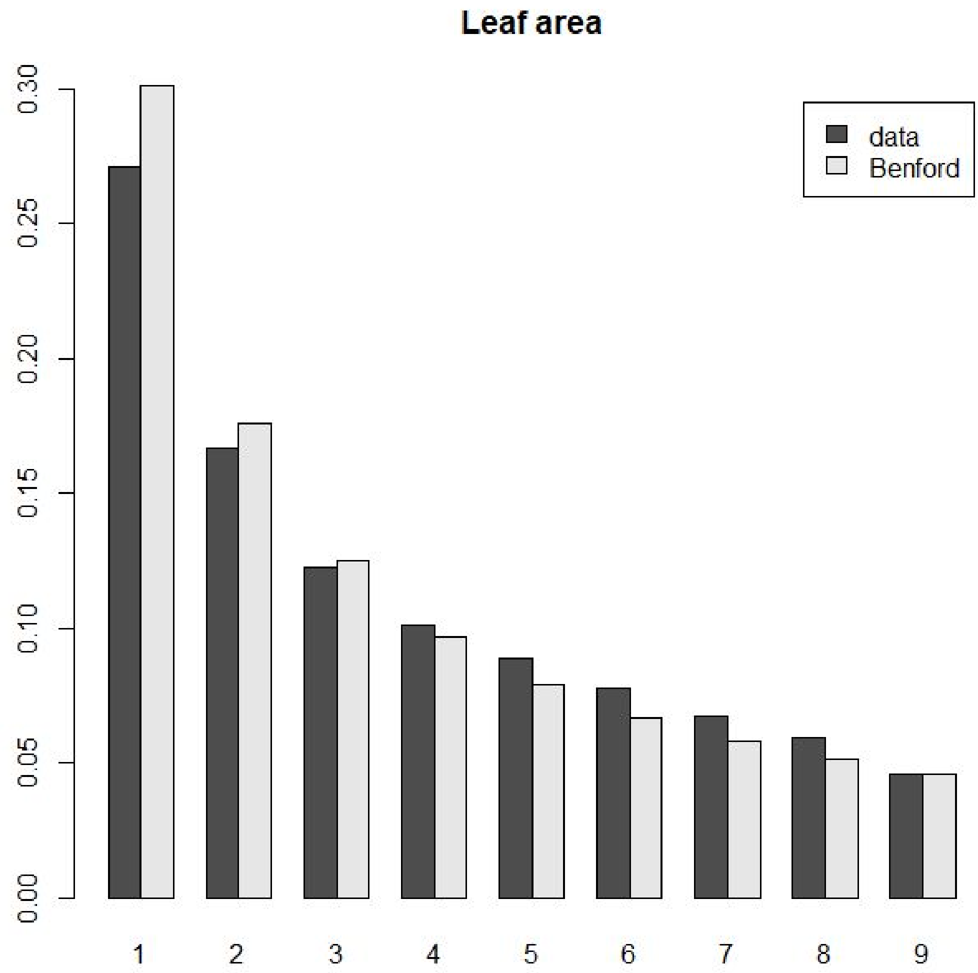
What to do

- Prepare the function `follow.benford (data)`, which will draw the barplot with frequencies of leading significant digits in the argument `data`, where `data` is a vector of numbers (either integers or real numbers, including negative values). The same barplot will also include a comparison with empirical (expected) values derived from Benford's law (see below). From each number in `data`, extract the leading (leftmost) digit (not zero, not a decimal point, and not a minus sign!).
- Count the frequency of individual digits (1-9, but not zero), and calculate the expected frequency of individual digits if they follow Benford's law: $\mathrm{P}\ (d) = \log_{10}\ ((d{+}1)/d)$, where d is the digit (e.g. 1), and $\mathrm{P}(d)$ is the probability that this digit will be the leading digit in the dataset (see Wikipedia). Simple calculation shows that $\mathrm{P}(1) = 0.301$, $\mathrm{P}(2) = 0.176$, $\mathrm{P}(3) = 0.125$, … $\mathrm{P}(9) = 0.046$.
- Test the function on the following two datasets (see examples of codes and results below): population sizes of 233 countries in the World (data from Wikipedia), and measured leaf areas for 3427 leaves from different habitats (data from Blonder et al. (2016), available from the Dataset website of Brian Enquist). Note that below I prepared both datasets for easy download from Gist on Github.

```
pop <- readr::read_delim ('https://gist.githubusercontent.com/zdealveindy/87f6ac90e989a1dc4c6b41f
follow.benford (pop$Population)
title (main = 'Population of countries in the World')
```



```
data.LA <- readr::read_delim ('https://gist.githubusercontent.com/zdealveindy/9efbebf002d3cccc752
follow.benford (data.LA$Wet.area.cm2)
title (main = 'Leaf area')
```


Leaf area
data
Benford
0.00
0.05
0.10
0.15
0.20
0.25
0.30
1
2
3
4
5
6
7
8
9