Data Science and Statistical Computing

Assignment 1

Due Monday 28th October 2024 at noon in Gradescope

Background

Benford's law is a purely phenomenological law proposed in 1938, when Frank Benford observed that the leading digit in a lot of real world settings is far from uniform (so you don't tend to see as many 9's as 8's as 7's etc as the first digit of a number). This may seem surprising, but was even first observed in the tables that were used to calculate logarithms before the advent of calculators/computers.

A set of leading digits is said to follow Benford's law if the probability that the leading digit is k is:

$$\mathbb{P}(k) = \log_{10}\left(1 + \frac{1}{k}\right), \quad k \in \{1, \ldots, 9\}$$

Note, zero is never a leading digit as it is just suppressed. Also note, $\mathbb{P}(k)$ clearly defines a valid probability mass function, since $\sum_{k=1}^{9} \mathbb{P}(k) = 1$.

Question

You are approached by a large bank who want your expertise to do an analysis of Google's corporate financial accounts. That is, the bank would like your help to test the hypothesis:

$$H_0 : \text{leading digits in Google's accounts obey Benford's law}$$
$$H_1 : \text{leading digits deviate from Benford's law}$$

This is the kind of hypothesis used by forensic accountants to search for potentially fraudulent financial filings. If H_0 is rejected, there may be evidence of fraud (at least, enough evidence to warrant careful scrutiny of the financial data), because humans are very bad at faking real data.

Below is the summary table of selected financial data from Google's 10-K SEC filing in the US[1].

	Year Ended December 31,				
	2015	2016	2017	2018	2019
	(in millions, except per share amounts)				
Consolidated Statements of Income Data:					
Revenues	$ 74,989	$ 90,272	$ 110,855	$ 136,819	$ 161,857
Income from operations	$ 19,360	$ 23,737	$ 26,178	$ 27,524	$ 34,231
Net income	$ 16,348	$ 19,478	$ 12,662	$ 30,736	$ 34,343

The leading digits from the 15 observed values above provide our Google SEC filing observations x_i to investigate the bank's hypothesis:

```
7, 9, 1, 1, 1
1, 2, 2, 2, 3
1, 1, 1, 3, 3
```

Let us denote these leading digits by $x_1, \ldots, x_{15}$.

[1] Google file under their parent company, called Alphabet Inc. ... see https://abc.xyz/investor/static/pdf/20200204_alphabet_10K.pdf?cache=cdd6dbf, page 26 if you are interested where this came from

1. Calculate $E_k, k \in \{1, \ldots, 9\}$, the expected number of leading digits being k in a data set of size 15, if it *does* obey Benford's law (ie when the null is true), assuming independence.

   (eg E_1 is the expected number of leading 1's in 15 observations, E_2 is the expected number of leading 2's in 15 observations, etc)

   Hint: the number of times each leading digit $k \in \{1, \ldots, 9\}$ appears in a sample of size 15 from $\mathbb{P}(k)$ will have a *Multinomial* distribution. This was mentioned in first year, but not in detail. See here for the expectation.

2. Let $O_k = \sum_{i=1}^{15} \mathbb{1}\{x_i = k\}$, be the number of leading digits k observed in a sample of size 15. Calculate $O_k, k \in \{1, \ldots, 9\}$, for the Google SEC filing data above.

3. The bank would like you to use the test statistic:

$$T = \sum_{k=1}^{9} (O_k - E_k)^2$$

   For this test statistic, will the p-value be $\mathbb{P}(T \geq t_{\text{obs}} \mid H_0 \text{ true})$, $\mathbb{P}(T \leq t_{\text{obs}} \mid H_0 \text{ true})$, or two sided?

   Hint: what range of values could T take? Are values at the upper, lower, or both ends of that range "extreme" compared to what happens when the null is true? The answer to that tells you correspondingly which probability statement is correct.

4. What is the observed test statistic, t_{obs}, for Google's SEC filing data above?

5. Your friend helps by doing a Monte Carlo simulation of 15 leading digits following Benford's law, and computes the test statistic above, 100 times in total. Below are the 100 simulated test statistics (which have been sorted from smallest to largest and put in a 20×5 grid for your convenience).

   Using these values, write down the Monte Carlo p-value estimate for the hypothesis test above, and state whether you would raise concerns over Google's financial accounts at any reasonable significance level?

```
 1.54  1.66  2.70  2.83  2.89
 3.02  3.07  3.16  3.71  4.44
 4.78  4.92  5.00  5.01  5.08
 5.12  5.35  5.38  5.48  5.56
 5.58  5.62  5.73  5.75  5.77
 6.05  6.07  6.11  6.64  6.69
 6.75  6.78  6.85  6.96  7.04
 7.24  7.47  7.92  8.05  8.34
 8.49  8.65  8.80  9.06  9.50
 9.57  9.85 10.01 10.02 10.34
10.36 10.44 10.60 10.70 11.02
11.58 12.11 12.14 12.17 12.18
12.23 12.62 12.91 12.97 13.01
13.27 13.38 13.58 14.61 14.74
14.86 15.12 15.31 15.31 15.34
15.83 16.08 16.71 16.96 16.99
17.49 18.34 18.41 18.52 18.62
18.98 19.48 19.70 20.87 21.39
22.35 22.64 23.35 23.89 24.77
24.82 27.35 29.92 30.08 40.87
```