Data Science and Statistical Computing

Assignment 1 Solutions

Q1

> 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.

Since this is a simple multinomial setting, we have that:

$$E_k = n\,\mathbb{P}(K = k) = 15 \log_{10}\left(1 + \frac{1}{k}\right)$$

This gives:

```
k  E_k
1 4.52
2 2.64
3 1.87
4 1.45
5 1.19
6 1.00
7 0.87
8 0.77
9 0.69
```

Q2

> 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.

```
k O_k
1   7
2   3
3   3
4   0
5   0
6   0
7   1
8   0
9   1
```

Q3

> 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.

In this case, when the null is true then O_k should be close to E_k. Of course, $O_k \neq E_k$ because $O_k \in \{0, 1, \ldots, 15\}$, so T cannot be zero, but smaller and positive is clearly indicative that we have observed something close to the null. Additionally, the test statistic is non-negative.

On the other hand, if the data does not follow the null then O_k will be further from E_k, so T will be large and positive.

Therefore, the p-value will be given by $\mathbb{P}(T \geq t_{\text{obs}} \mid H_0 \text{ true})$.

Q4

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

```
12.7814
```

Q5

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
```

We can see that t_{obs} sits between the simulated values 12.62 and 12.91. There are a total of $3+7\times5 = 38$ of the 100 simulations which are greater than or equal to this, so $p = 0.38$.

There is no reasonable significance level which would consider this to be a significant result. Therefore, there is no evidence within these summary accounts to cause any concerns of deviation from Benford's Law.