Exact Asymptotic Goodness-of-Fit Testing For Discrete Circular Data, With Applications

David E. Giles
1. Background

2. Motivation

3. Existing results

4. Main contribution(s)

5. Overview – main theoretical results

6. Applications

7. Conclusions
1. Background

- Construction of goodness-of-fit tests when data are distributed on the circle (sphere, hypersphere) is an important statistical problem.

- Tests that have been proposed for continuous data include Kuiper’s (1959) $V_N$ test and Watson’s (1961) $U_N^2$ test.

- These tests are of the Kolmogorov-Smirnov type, being based on the empirical distribution function.

- Rely on the Glivenko-Cantelli Theorem.
Two general types of test, based on:

1. Maximum “gap” between theoretical and empirical c.d.f.’s.
   *e.g.*, Kolmogorov; Kuiper; Watson; Lilliefors.

2. “Area” between the theoretical and empirical c.d.f.’s.
   *e.g.*, Anderson-Darling; Cramer – von Mises.

- Modify tests if the data are “circular”:
2. Motivation

- Benford’s law(s). Wanted to test with discrete, circular, data.

- Benford (1938) re-discovered Newcomb’s (1881) observation that the first significant digit \(d_1\) of certain naturally occurring numbers follows the distribution given by

\[
p_i = \Pr[d_1 = i] = \log_{10}[1 + (1/i)] \; ; i = 1, 2, \ldots, 9.\]

- Joint distributions for first two and first three digits \(d_1, d_2\) and \(d_3\):

\[
p_{ij} = \Pr[d_1 = i, d_2 = j] = \log_{10}[1 + 1/(10i + j)] \; ; i, j = 10, 11, \ldots, 99\]
\[ p_{ijk} = \Pr[d_1 = i, d_2 = j, d_3 = k] = \log_{10}[1 + 1/(100i + 10j + k)]; \]
\[ i, j, k = 100, 101, \ldots, 999. \]

- Marginal distributions for \( d_2 \) and \( d_3 \) are

\[ p_i = \Pr[d_2 = i] = \sum_{l=1}^{9} \log_{10}[1 + 1/(10l + i)]; \quad i = 0, 1, \ldots, 9 \]

\[ p_i = \Pr[d_3 = i] = \sum_{l=1}^{9} \sum_{m=0}^{9} \log_{10}[1 + 1/(100l + 10m + i)]; \quad i = 0, 1, \ldots, 9. \]

- Numerous applications and examples, including auditing, hydrology, physics, survey data, eBaY auction prices (Giles, 2007), psychological barriers in market prices (Lu and Giles, 2010).
Benford's First-Digit Law
Benford's Second-Digit & Third-Digit Laws
3. Existing results

- K-S type tests for discrete data have received far less attention in the literature.

- Why is this?

- K-S statistics are distribution-free in the continuous case, but generally not when the data are discrete.

- Usual tests have to be modified.

- \( H_0: \) The data follow a discrete circular distribution, \( F \), defined by the probabilities \( \{p_i\} \). \( H_1: \) \( H_0 \) is not true. Sample of \( N \) observations.
• Let \( \{r_i\}_{i=1}^n \) denote the sample frequencies, such that \( \sum_{i=1}^n r_i = N \).

• For this general problem, Freedman (1981) proposed a modified version of Watson’s \( U^2_N \) test for use with discrete data.

• He provided Monte Carlo evidence that this test out-performs Kuiper’s (1962) modified test for the discrete case.

• Freedman’s test statistic is:

\[
U^*_{N} = (N / n)\left[ \sum_{j=1}^{n-1} S_j^2 - \left( \sum_{j=1}^{n-1} S_j \right)^2 / n \right],
\]

where

\[
S_j = \sum_{i=1}^{j} (r_i / N - p_i) ; \quad j = 1, 2, \ldots, n.
\]
Asymptotic null distribution of the test statistic is a weighted sum of \((n - 1)\) independent chi-squared variates, each with one degree of freedom, and with weights which are the eigenvalues of the matrix whose \((i, j)^{th}\) element is 
\[
\left(\frac{p_i}{n^2}\right)\left(\{n - \max(i, j)\} \min(i, j) - \sum_{k=1}^{n-1} p_k \{n - \max(i, j)\} \min(j, k)\right).
\]

Freedman expressed the first four moments of the asymptotic distribution of the test statistic under \(H_0\) as functions of these eigenvalues.

Used these moments to approximate the quantiles of the asymptotic distribution by fitting Pearson curves.

*He only considered the case where the population distribution is uniform multinomial.*
4. **Main contribution(s)**

*In fact.....................*

- Asymptotic null distribution of $U^*_N$ can be obtained directly and *without any approximations* by using standard computational methods.

- Specifically, we can use those suggested by Imhof (1961), Davies (1973, 1980) and others, to invert the characteristic function for statistics which are weighted sums of chi-squared variates.

- There is no need to resort to approximations, curve fitting or simulation methods.

- Davies’ algorithm readily available - *e.g.*, in SHAZAM.
This paper demonstrates how to obtain accurate quantiles for $U^*_{N^2}$.

Quantiles are tabulated for discrete uniform, Benford (1st, 2nd, 3rd) and beta-binomial null distributions.

Several illustrative applications are provided.

Why not simulate the quantiles?

Very inaccurate in tails.

Need to use “rare event” methods such as generalized splitting and hit-and-run sampling (e.g., Grace and Wood, 2012).
5. **Main theoretical results**

- Quantiles for $U_{N}^{*2}$ asymptotic null distribution, $H_0$: Discrete uniform.
- Quantiles for $U_{N}^{*2}$ asymptotic null distribution, $H_0$: Benford 1\textsuperscript{st} digit.
- Quantiles for $U_{N}^{*2}$ asymptotic null distribution, $H_0$: Benford 2\textsuperscript{nd} digit.
- Quantiles for $U_{N}^{*2}$ asymptotic null distribution, $H_0$: Benford 3\textsuperscript{rd} digit.
- Quantiles for $U_{N}^{*2}$ asymptotic null distribution, $H_0$: Beta-binomial.
- Power of $U_{N}^{*2}$ test.
Figure 1: Exact Asymptotic Distribution of Freeman’s Statistic for the Uniform Discrete Distribution Under the Null Hypothesis
Figure 2: Exact Asymptotic Distributions of Freeman's Statistic for Benford's Distributions for First and Second Digits Under the Null Hypothesis

Note: Distributions for second and third digits are visually indistinguishable
Figure 3: Exact Asymptotic Distribution of Freeman's Statistic for the Beta-Binomial Distribution With n = 12 Under the Null Hypothesis

- $\alpha = 0.2; \beta = 0.25$
- $\alpha = 0.7; \beta = 0.2$
- $\alpha = 2.0; \beta = 2.0$
- $\alpha = 600; \beta = 400$
6. **Applications**

**Canadian births**

- Demographic literature suggests there will non-uniformity due to seasonal effects.
- Data are circular and discrete \((n = 12)\)
- Reject uniformity for Canada as a whole and for all regions except PEI, NWT, YT and NU.
Table 4. Canadian live births, 2008: relative frequency distribution (%)

<table>
<thead>
<tr>
<th>Month:</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>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>NL</td>
<td>7.4</td>
<td>7.4</td>
<td>8.4</td>
<td>7.8</td>
<td>8.8</td>
<td>7.8</td>
<td>8.8</td>
<td>9.7</td>
<td>9.6</td>
<td>8.9</td>
<td>7.8</td>
<td>7.6</td>
</tr>
<tr>
<td>PEI</td>
<td>7.3</td>
<td>9.0</td>
<td>9.0</td>
<td>7.6</td>
<td>8.6</td>
<td>8.2</td>
<td>9.6</td>
<td>7.6</td>
<td>8.3</td>
<td>8.4</td>
<td>8.7</td>
<td>7.6</td>
</tr>
<tr>
<td>NS</td>
<td>8.3</td>
<td>8.1</td>
<td>8.0</td>
<td>8.1</td>
<td>8.5</td>
<td>8.4</td>
<td>9.5</td>
<td>8.5</td>
<td>8.9</td>
<td>8.5</td>
<td>7.6</td>
<td>7.5</td>
</tr>
<tr>
<td>NB</td>
<td>8.0</td>
<td>7.7</td>
<td>8.3</td>
<td>7.7</td>
<td>8.4</td>
<td>8.5</td>
<td>8.7</td>
<td>9.3</td>
<td>9.0</td>
<td>8.5</td>
<td>7.9</td>
<td>7.9</td>
</tr>
<tr>
<td>QC</td>
<td>7.7</td>
<td>7.6</td>
<td>8.2</td>
<td>8.2</td>
<td>8.5</td>
<td>8.2</td>
<td>9.2</td>
<td>8.7</td>
<td>9.0</td>
<td>8.9</td>
<td>7.9</td>
<td>7.9</td>
</tr>
<tr>
<td>ON</td>
<td>8.2</td>
<td>7.8</td>
<td>8.2</td>
<td>8.4</td>
<td>8.6</td>
<td>8.4</td>
<td>8.8</td>
<td>8.6</td>
<td>8.9</td>
<td>8.6</td>
<td>7.9</td>
<td>7.8</td>
</tr>
<tr>
<td>MB</td>
<td>8.2</td>
<td>7.6</td>
<td>7.9</td>
<td>8.1</td>
<td>8.7</td>
<td>8.3</td>
<td>9.0</td>
<td>8.8</td>
<td>8.9</td>
<td>9.0</td>
<td>7.5</td>
<td>8.0</td>
</tr>
<tr>
<td>SK</td>
<td>8.2</td>
<td>8.0</td>
<td>8.3</td>
<td>8.2</td>
<td>8.8</td>
<td>8.4</td>
<td>8.7</td>
<td>8.3</td>
<td>9.5</td>
<td>8.3</td>
<td>7.4</td>
<td>7.9</td>
</tr>
<tr>
<td>AB</td>
<td>8.0</td>
<td>7.6</td>
<td>8.2</td>
<td>8.4</td>
<td>8.6</td>
<td>8.7</td>
<td>8.9</td>
<td>8.9</td>
<td>8.6</td>
<td>8.4</td>
<td>7.6</td>
<td>8.1</td>
</tr>
<tr>
<td>BC</td>
<td>8.0</td>
<td>7.6</td>
<td>8.1</td>
<td>8.2</td>
<td>8.8</td>
<td>8.4</td>
<td>8.9</td>
<td>8.7</td>
<td>8.9</td>
<td>8.4</td>
<td>7.8</td>
<td>8.2</td>
</tr>
<tr>
<td>YT</td>
<td>6.2</td>
<td>7.8</td>
<td>9.1</td>
<td>6.4</td>
<td>10.2</td>
<td>6.7</td>
<td>5.9</td>
<td>9.4</td>
<td>10.7</td>
<td>9.1</td>
<td>8.3</td>
<td>10.2</td>
</tr>
<tr>
<td>NWT</td>
<td>8.7</td>
<td>7.2</td>
<td>8.6</td>
<td>8.5</td>
<td>9.4</td>
<td>7.8</td>
<td>8.2</td>
<td>10.3</td>
<td>7.9</td>
<td>7.9</td>
<td>8.5</td>
<td>7.1</td>
</tr>
<tr>
<td>NU</td>
<td>7.7</td>
<td>7.7</td>
<td>9.3</td>
<td>8.9</td>
<td>9.2</td>
<td>9.6</td>
<td>8.8</td>
<td>8.3</td>
<td>8.7</td>
<td>6.7</td>
<td>7.5</td>
<td>7.6</td>
</tr>
<tr>
<td>CAN</td>
<td>8.0</td>
<td>7.7</td>
<td>8.2</td>
<td>8.3</td>
<td>8.6</td>
<td>8.4</td>
<td>8.9</td>
<td>8.7</td>
<td>8.9</td>
<td>8.6</td>
<td>7.8</td>
<td>7.9</td>
</tr>
</tbody>
</table>
## Table 5. Values of $U^*_{N}$. $H_0$: Canadian birth months follow uniform discrete distribution

<table>
<thead>
<tr>
<th>Province/Territory</th>
<th>$N$</th>
<th>$U^*_{N}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>NL</td>
<td>4,898</td>
<td>0.771</td>
</tr>
<tr>
<td>PEI</td>
<td>1,483</td>
<td>0.038</td>
</tr>
<tr>
<td>NS</td>
<td>9,188</td>
<td>0.528</td>
</tr>
<tr>
<td>NB</td>
<td>7,402</td>
<td>0.490</td>
</tr>
<tr>
<td>QC</td>
<td>87,870</td>
<td>6.340</td>
</tr>
<tr>
<td>ON</td>
<td>140,791</td>
<td>5.681</td>
</tr>
<tr>
<td>MB</td>
<td>15,485</td>
<td>0.994</td>
</tr>
<tr>
<td>SK</td>
<td>13,737</td>
<td>0.552</td>
</tr>
<tr>
<td>AB</td>
<td>50,856</td>
<td>2.856</td>
</tr>
<tr>
<td>BC</td>
<td>44,276</td>
<td>2.093</td>
</tr>
<tr>
<td>YT</td>
<td>373</td>
<td>0.089</td>
</tr>
<tr>
<td>NWT</td>
<td>721</td>
<td>0.052</td>
</tr>
<tr>
<td>NU</td>
<td>805</td>
<td>0.168</td>
</tr>
<tr>
<td>CANADA</td>
<td>377,886</td>
<td>18.146</td>
</tr>
</tbody>
</table>
Fibonacci numbers and factorials

1. Fibonacci numbers should follow Benford’s 1st - digit law (Duncan, 1969; Washington, 1981; Canessa, 2003). *Not previously tested.*
   - Consider up to the first $N = 20,000$ numbers; $n = 9$.
   - Reject uniformity, $N \geq 50$. Cannot reject Benford’s law, for any $N$.

2. Factorials and binomial coefficients should follow Benford’s 1st - digit law (Sarkar, 1973). *Not previously tested.*
   - Maximum $N = 170$, $n = 9$.
   - Reject uniformity for all $N$; cannot reject Benford’s law if $N > 50$. 
Benford's First-Digit Law
Table 8 (a). Values of $U^*_{N}$. $H_0$: Fibonacci first digits follow uniform discrete distribution; or $H_0$: Fibonacci first digits follow Benford’s distribution

<table>
<thead>
<tr>
<th>$N$</th>
<th>$U^*_{N}$</th>
<th>$U^*_{N}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$H_0$: Uniform discrete</td>
<td>$H_0$: Benford</td>
</tr>
<tr>
<td>50</td>
<td>0.42831</td>
<td>0.00486</td>
</tr>
<tr>
<td>100</td>
<td>0.79613</td>
<td>0.00342</td>
</tr>
<tr>
<td>500</td>
<td>3.84342</td>
<td>0.00063</td>
</tr>
<tr>
<td>1000</td>
<td>7.71638</td>
<td>0.00042</td>
</tr>
<tr>
<td>2000</td>
<td>13.35437</td>
<td>0.00021</td>
</tr>
<tr>
<td>5000</td>
<td>38.44199</td>
<td>0.00012</td>
</tr>
<tr>
<td>10000</td>
<td>76.84573</td>
<td>0.00007</td>
</tr>
<tr>
<td>20000</td>
<td>153.54990</td>
<td>0.00003</td>
</tr>
</tbody>
</table>
(b). Values of $U_{N}^{*2}$. $H_0$: Factorials first digits follow uniform discrete distribution; or $H_0$: Factorials first digits follow Benford’s distribution

<table>
<thead>
<tr>
<th>$N$</th>
<th>$U_{N}^{*2}$</th>
<th>$H_0$: Uniform discrete</th>
<th>$H_0$: Benford</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>1.16915</td>
<td></td>
<td>0.27684</td>
</tr>
<tr>
<td>100</td>
<td>1.47179</td>
<td></td>
<td>0.08815</td>
</tr>
<tr>
<td>170</td>
<td>1.56025</td>
<td></td>
<td>0.04822</td>
</tr>
</tbody>
</table>

(10% = 0.154; 5% = 0.191; 1% = 0.276)

(10% = 0.143; 5% = 0.179; 1% = 0.263)
eBay auction prices

- Price data exhibit circularity. Consider two prices such as $99.99 and $100. Their first significant digits are as far apart as is possible, yet the associated prices are extremely close.

- Giles (2007) considered all of the 1,161 successful auctions for tickets for professional football games in the “event tickets” category on eBay for the period 25 November to 3 December, 2004, excluding auctions ending with the “Buy-it-Now” option, and all Dutch auctions.

- The winning bids should satisfy Benford’s Law if they are “naturally occurring” numbers, as should be the case if there were no collusion among bidders and no “shilling” by sellers in this market.
Auction Price Data - Benford's First-Digit Law
Auction Price Data - Benford's Second-Digit Law
Auction Price Data - Benford's Third-Digit Law
- Uniformity is strongly rejected (against non-uniformity) for the first and third digits, and for the second digit if $N \geq 250$.
- At the 5% significance level, Benford’s Law for the third digit is unambiguously rejected (against the non-Benford alternative), and the first digit and second digit laws are also rejected for $N > 100$.
- Suggests that these auction prices are not “naturally occurring numbers”.
- Could be evidence of collusion or “shilling”.
Table 9. Values of $U_N^{*2}$. $H_0$: Football ticket price digits follow uniform discrete distribution; or $H_0$: Football ticket price digits follow Benford’s distribution

<table>
<thead>
<tr>
<th>$N$</th>
<th>Digit 1</th>
<th>Digit 2</th>
<th>Digit 3</th>
<th>$U_N^{*2}$ Uniform discrete</th>
<th>Digit 1</th>
<th>Digit 2</th>
<th>Digit 3</th>
<th>$U_N^{*2}$ Benford</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0.4574</td>
<td>0.1242</td>
<td>0.3952</td>
<td>0.0463</td>
<td>0.1094</td>
<td>0.3883</td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>1.1306</td>
<td>0.0476</td>
<td>0.2490</td>
<td>0.0778</td>
<td>0.0195</td>
<td>0.2407</td>
<td></td>
<td></td>
</tr>
<tr>
<td>250</td>
<td>3.3508</td>
<td>0.7566</td>
<td>2.7800</td>
<td>0.2673</td>
<td>0.5128</td>
<td>2.7390</td>
<td></td>
<td></td>
</tr>
<tr>
<td>500</td>
<td>5.6113</td>
<td>1.1440</td>
<td>4.8389</td>
<td>0.2539</td>
<td>0.6876</td>
<td>4.7680</td>
<td></td>
<td></td>
</tr>
<tr>
<td>750</td>
<td>8.3334</td>
<td>1.4935</td>
<td>6.9105</td>
<td>0.3210</td>
<td>0.8473</td>
<td>6.7987</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>10.6368</td>
<td>2.1118</td>
<td>9.2640</td>
<td>0.2919</td>
<td>1.2482</td>
<td>9.1235</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1161</td>
<td>11.7730</td>
<td>2.4803</td>
<td>11.1671</td>
<td>0.2258</td>
<td>1.4664</td>
<td>10.9962</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(5%: 0.191373 0.17878 0.19016 0.19052)
7. **Conclusions**

- Need to take special care if using EDF tests when data are discrete.
- Additional issues if data are also “circular”.
- Freedman’s modification of Watson’s $U^*_N$ test is recommended.
- Quantiles of null asymptotic distribution can be computed exactly.
- Quantiles presented for Uniform, Benford (1, 2, 3) and Beta-Binomial null distributions.
- Various illustrative applications.
  - The $U^*_N$ test has 100% power, for $N \geq 150$ when testing nulls of either Benford’s distribution or beta-binomial distribution, against uniform alternatives.