University of Victoria
Department of Economics

Econometrics Working Paper EWP0607

ISSN 1485-6441

THE EXACT ASYMPTOTIC DISTRIBUTION FUNCTION OF WATSON'S U_N^2 FOR TESTING GOODNESS-OF-FIT WITH CIRCULAR DISCRETE DATA*

David E. Giles

Department of Economics, University of Victoria
Victoria, B.C., Canada V8W 2Y2

November, 2006

Abstract

We show that the full asymptotic distribution for Watson's U_N^2 statistic, modified for discrete data, can be computed by standard methods. Previous approximate percentiles for the uniform multinomial case are found to be accurate. More extensive percentiles are presented for this distribution, and for the distribution associated with "Benford's Law".

* We are grateful to Sophie Hesling for her assistance with data collection.

Keywords: Distributions on the circle; Goodness-of-fit; Watson's U_N^2; discrete data; Benford's Law

JEL Classifications: C12; C16; C46

Author Contact:
David E. Giles, Dept. of Economics, University of Victoria, P.O. Box 1700, STN CSC, Victoria, B.C., Canada V8W 2Y2; e-mail: dgiles@uvic.ca; Phone: (250) 721-8540; FAX: (250) 721-6214

1. INTRODUCTION

Testing for goodness-of-fit when the data are distributed on the circle (or more generally the sphere) is an important statistical problem. An excellent discussion is provided, for example, by Watson (1973). Among the tests that have been proposed for continuous data are Kuiper's (1959) V_N test and Watson's (1961) U_N^2 test. Detailed significance points are provided by Stephens (1964, 1965). Testing for goodness-of-fit on the circle in the case of discrete data has received far less attention in the literature.

Suppose that we have a discrete distribution defined by the probabilities $\{p_i\}_{i=i}^n$, and let $\{r_i\}_{i=i}^n$ denote the corresponding sample frequencies, such that $\sum_{i=1}^{n} r_i = N$. Freedman (1981) proposes the following modified version of Watson's U_N^2 statistic for discrete distributions:

$$U_N^2 = (N/n)\left[\sum_{j=1}^{n-1} S_j^2 - \left(\sum_{j=1}^{n-1} S_j\right)^2 / n\right] \quad , \tag{1}$$

where

$$S_j = \sum_{i=1}^{j} (r_i / N - p_i) \quad ; \quad j = 1, 2, \ldots, n.$$

He also shows that the asymptotic distribution of the statistic in (1) 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 $(i, j)^{\text{th}}$ element is

$$(p_i / n^2)\left(\{n - \max(i, j)\}\min(i, j) - \sum_{k=1}^{n-1} p_k \{n - \max(i, j)\}\min(j, k)\right) \quad .$$

Freedman expresses the first four moments of the asymptotic distribution as functions of these eigenvalues, and uses these moments to approximate percentiles of the asymptotic distribution by fitting Pearson curves. The quality of this approximation is confirmed by Monte Carlo methods for the case where the population distribution is uniform multinomial. In fact, the complete asymptotic distribution can be obtained directly by using standard computational methods, such as those suggested by Imhof (1961), Davies (1973, 1980) and others.

2. ASYMPTOTIC DISTRIBUTION

An advantage of Davies' algorithm is its accuracy and the fact that the latter can be controlled by the user. In what follows we use Davies' own double-precision FORTRAN code that has been incorporated into the SHAZAM econometrics package (Whistler *et al*., 2004). The integration error bound and maximum number of integration terms for the inversion of the characteristic function were set to 10^{-6} and 1000 respectively. The calculations were undertaken on a PC with an Intel Pentium 3.00 GHz processor, running Windows XP Pro.

Figure 1 shows the full asymptotic distribution function of U_N^2 for the uniform multinomial case, for selected values of *n*. Table 1 provides percentiles for a wider range of *n*, and compares these with Freeman's approximate percentiles as appropriate. The case of $n = 12$ is of interest when testing for seasonal incidence with monthly data. Freedman's Pearson curves provide more (less) accurate upper (lower) percentiles than those obtained from Monte Carlo simulation.

As a second example, consider the discrete distribution usually referred to as "Benford's Law". Benford (1938) re-discovered Newcomb's (1881) observation that the first significant digit (*d*) of certain naturally occurring numbers follows the distribution given by

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

Applications of this distribution occur in the auditing of financial data (*e.g.*, Geyer and Williamson, 2004), and in testing for collusion or "shilling" in auctions (*e.g*., Giles, 2006). Figure 2 depicts the distribution function for the discrete version of U_N^2 for Benford's distribution, and Table 2 provides a range of associated percentiles.

3. APPLICATIONS

Canessa (2003) has proposed a general statistical thermodynamic theory that explains, *inter alia*, why Fibonacci sequences should obey Benford's Law. The distribution of the first digits of the first *N* of the Fibonacci series, {1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144, ……} has been tested against Benford's Law, for various choices of $N \leq 1{,}476$ (the upper bound being determined by the largest Fibonacci number storable in an Excel worksheet). The results, in Table 3, indicate a

clear rejection of uniformity (using the percentiles for $n = 9$ in Table 1) and equally clear support for Benford's Law.

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 (2006) 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 would be the case if there were no collusion among bidders and no "shilling" by sellers in this market. Table 4 reports the results of testing these first digits against the uniform multinomial and Benford hypotheses. Uniformity is again strongly rejected. Benford's Law cannot be rejected for the first 100 winning bids, but it is rejected at the 2.5% significance level for $N > 100$. In contrast, with $N = 1{,}161$, Giles (2006) marginally fails to reject Benford's Law when (wrongly) applying Kuiper's (1959) V_N test for continuous data to these integer values.

Table 1. *Percentiles of the asymptotic distribution function of U_N^2 for the uniform multinomial distribution*

n	Lower Percentiles					Upper Percentiles				
	1	2.5	5	10	25	75	90	95	97.5	99
3	0.0007	0.0019	0.0038	0.0078	0.0213	0.1027	0.1706	0.2219	0.2733	0.3411
4	0.0029	0.0054	0.0088	0.0146	0.0305	0.1064	0.1649	0.2086	0.2521	0.3094
5	0.0054	0.0088	0.0129	0.0194	0.0357	0.1070	0.1609	0.2012	0.2414	0.2944
6	0.0076	0.0115	0.0160	0.0228	0.0389	0.1069	0.1583	0.1969	0.2354	0.2864
7	0.0095	0.0137	0.0183	0.0251	0.0409	0.1066	0.1567	0.1943	0.2319	0.2815
8	0.0111	0.0154	0.0200	0.0267	0.0422	0.1064	0.1556	0.1926	0.2295	0.2784
9	0.0124	0.0167	0.0213	0.0279	0.0431	0.1062	0.1548	0.1914	0.2280	0.2763
10	0.0134	0.0177	0.0223	0.0288	0.0437	0.1060	0.1542	0.1905	0.2268	0.2748
15	0.0164	0.0206	0.0249	0.0311	0.0452	0.1056	0.1529	0.1885	0.2241	0.2729
20	0.0177	0.0217	0.0260	0.0320	0.0457	0.1055	0.1524	0.1878	0.2232	0.2700
30	0.0188	0.0226	0.0267	0.0326	0.0461	0.1053	0.1520	0.1873	0.2226	0.2691
12	0.0149	0.0192	0.0237	0.0301	0.0446	0.1058	0.1535	0.1894	0.2253	0.2729
	(0.0195)	(0.0218)	(0.0248)	(0.0299)	(0.0435)	(0.106)	(0.154)	(0.189)	(0.225)	(0.272)
	[0.015]	[0.019]	[0.024]	[0.030]	[0.045]	[0.107]	[0.155]	[0.191]	[0.224]	[0.264]

Note: For $n = 12$, figures in parentheses are Freedman's (1981) Pearson curve approximations, and those in square brackets are his Monte Carlo estimates. More extensive results to more decimal places are available from the author.

Table 2. *Percentiles of the asymptotic distribution function of U_N^2 for Benford's distribution*

Lower Percentiles		Upper Percentiles	
1	0.01024	75	0.09651
2.5	0.01392	90	0.14313
5	0.01794	95	0.17878
10	0.02379	97.5	0.21485
25	0.03744	99	0.26319

Table 3. *Values of U_N^2 when testing Fibonacci first digits against uniform multinomial and Benford's distributions*

N	U_N^2	
	Uniform	Benford
50	0.3345	0.0049
100	0.5678	0.0034
250	1.4246	0.0018
500	2.7328	0.0007
750	3.5323	0.0299
1000	4.8953	0.0216
1200	6.0650	0.0156
1476	7.5542	0.0127

Table 4. *Values of U_N^2 when testing football auction price first digits against uniform multinomial and Benford's distributions*

N	U_N^2	
	Uniform	Benford
50	0.4513	0.0463
100	0.6938	0.0778
250	1.5283	0.2218
500	3.6757	0.2713
750	4.9501	0.3203
1000	6.6153	0.3065
1161	7.3380	0.2336

Figure 1: Exact Asmptotic Distribution of Watson's Statistic for Uniform Multinomial Population

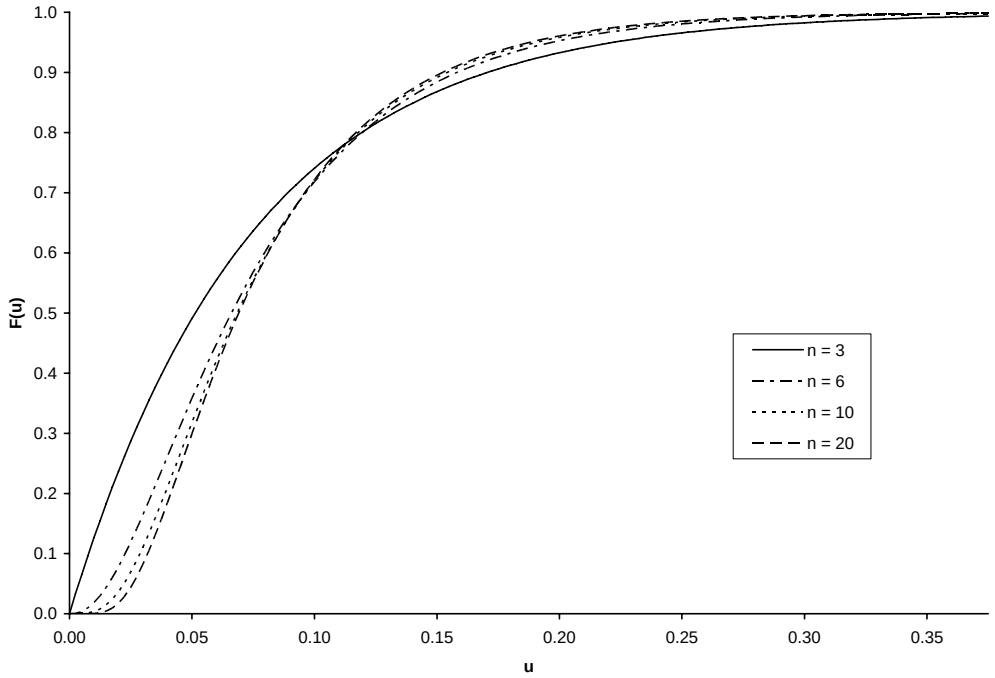
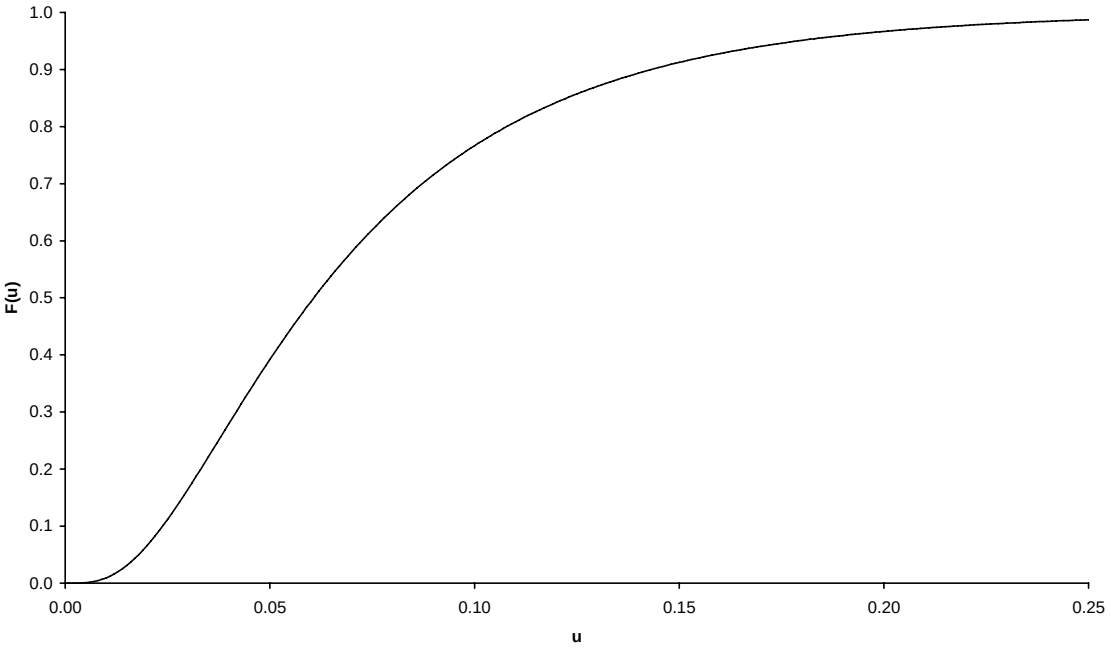


Figure 2: Exact Asymptotic Distribution of Watson's Statistic for Population Following Benford's Distribution



REFERENCES

BENFORD F. (1938). The law of anomalous numbers. *Proceedings of the American Philosophical Society* **78**, 551-72.

CANESSA, E. (2003). Theory of analogous force in number sets. *Physica A* **328**, 44-52.

DAVIES, R. B. (1973). Numerical inversion of a characteristic function. *Biometika* **60**, 415-7.

DAVIES, R. B. (1980). The distribution of a linear combination of χ^2 random variables, algorithm AS 155.*Applied Statistics* **29**, 323-33.

FREEDMAN, L. S. (1981). Watson's U_N^2 statistic for a discrete distribution. *Biometrika* **68**, 708-11.

GEYER, C. L. & WILLIAMSON, P. P. (2004). Detecting fraud in data sets using Benford's Law. *Communications in Statistics B* **33**, 229-46.

GILES, D. E. A. (2006). Benford's Law and naturally occurring prices in certain eBay auctions. *Applied Economics Letters*, in press.

IMHOF, J. P. (1961). Computing the distribution of quadratic forms in normal variables. *Biometrika* **48**, 419-26.

KUIPER, N. H. (1959). Alternative proof of a theorem of Birnbaum and Pyke. *Annals of Mathematical Statistics*. **30**, 251-2.

NEWCOMB, S. (1881). Note on the frequency of use of the different digits in natural numbers. *American Journal of Mathematics* **4**, 39-40.

STEPHENS, M. A. (1964). The distribution of the goodness-of-fit statistic U_N^2, II. *Biometrika* **51**, 393-7.

STEPHENS, M. A. (1965). The goodness-of-fit statistic, V_N : Distribution and significance points. *Biometrika* **52**, 309-21.

WATSON, G. S. (1961).Goodness-of-fit tests on a circle.I. *Biometrika* **48**, 109-14.

WATSON, G. S. (1973). *Distribution Theory for Tests Based on the Sample Distribution Function*. Society for Industrial and Applied Mathematics, Philadelphia, PA.

WHISTLER, D., WHITE, K. J., Wong, S. D. & BATES, D. (2004). *SHAZAM Econometrics Software: User's Reference Manual Version 10*. Northwest Econometrics, Vancouver BC.