Goodness-of-Fit Testing for the Zipf Distribution

David E. Giles

Department of Economics

University of Victoria

March 2024

Abstract

The Zipf distribution is a truncated power law that has been used widely in many disciplines. In this paper we develop a new formal test for the appropriateness of the Zipf distribution, based on the empirical distribution function, that allows for the discrete nature of the data. Exact quantiles for the null distribution of the test statistic are determined, and the asymptotic test is shown to have good power properties, even in small samples. Several illustrative applications of the test are provided.

Key Words Zipf distribution; empirical distribution function; hypothesis testing

Author contact

58 Rock Lake Court, RR1, Bancroft, Ontario K0L 1C0, CANADA; dgiles@uvic.ca ; +1-613-202 9501

(*The author is Professor Emeritus, Department of Economics, University of Victoria, CANADA*.)

1. Introduction

The Zipf distribution and "Zipf's law" (Zipf, 1949) have received considerable attention across a broad range of disciplines. An important aspect of this literature is the task of testing whether or not a set of data is consistent with this distribution. Although this issue has been addressed from different viewpoints by a number of authors, a formal goodness-of-fit test that takes proper account of the discrete nature of the associated data has received little or no attention to date. This paper fills that gap.

If a discrete random variable, *X*, follows a Zipf distribution, its p.m.f. is:

$$p(x) = Pr.[X = x] = x^{-s}/H_{n,s} \quad ; \quad x = 1, 2, 3, \dots, n \; ; s > 0 \tag{1}$$

where $H_{n,s} = \sum_{k=1}^{n} k^{-s}$ is the n^{th} generalized harmonic number, with exponent *s*. In other words, *X* follows a truncated, discrete, power law, with scaling parameter, *s*.

All of the integer-order moments of the Zipf distribution are finite, with the m^{th} moment about the origin being $\mu_m = (\frac{H_{n,s-m}}{H_{n,s}})$; m = 1, 2, …. Obviously, for some manipulations involving the distribution, a closed-form expression for $H_{n,s}$ would be useful, and Naldi (2015) offers various approximations to this.

The Zipf distribution is a right-truncated version of the Zeta distribution, with the latter's p.m.f. being:

$$p(x) = Pr.[X = x] = x^{-s}/\zeta(s) \quad ; \quad x = 1, 2, 3, \dots \dots \dots \; ; s > 1 \tag{2}$$

where $\zeta(s) = \sum_{k=1}^{\infty} k^{-s}$ is the Riemann zeta function, which has a simple pole at $s = 1$. Alternatively, the Zeta distribution can be viewed as a limiting case of the Zipf distribution. The r^{th} moment of (2) exists only if $s > r + 1$.

The so-called Zipf's law can be extracted from (1) when *s* = 1. Then, the probability that the event {*X* = *x*} occurs is inversely proportional to the value of *x*. This also implies that the frequency of occurrence, *f*, of the value *x* is inversely proportional its value. More generally, the Zipf distribution describes the situation where $f(x)x^s = constant$, where $s > 0$. This implies that a plot of $\log[f(x)]$ against $\log(x)$ will be a straight line, with slope equal to -*s*.

As noted above, there is a vast literature associated with Zipf's law, and with the discrete distribution in (1). References to this literature can be found in recent contributions, which include Boamah-Addo *et al*. (2022), Gabaix (1999, 2009, 2016), Gabaix and Ibragimov (2011), Giesen and Südekum (2011),

Hinloopen and van Marrewijk (2012), Naldi (2003, 2015), Newman (2005), Zanette (2006), and Zanette and Montemurro (2005), among others.

One part of this literature deals with the problem of examining whether or not a sample of data is generated by the Zipf distribution. For example, see Clauset *et al*. (2009), Schluter (2021), and Urzúa (2000). The objective of this paper is to provide an appropriate and formal test of this hypothesis. In the next section we discuss some of the issues associated with conventional goodness-of-fit tests when the data are discrete, rather than continuous in nature. An appropriate asymptotic test proposed by Freedman (1981), and modified computationally by Giles (2013), is applied to the Zipf distribution. Section 3 provides the exact asymptotic quantiles needed to apply the test in practice; and investigates the power of the test in finite samples against some relevant alternative distributions. Several applications using actual data are presented in section 4; and section 5 provides some concluding remarks.

2. Goodness-of-fit testing with discrete data

There is an extensive literature relating to goodness-of-fit tests based on the empirical distribution function (EDF), and useful discussions of this literature are provided by Stephens (1974) and D'Agostino and Stephens (1986), for example. Justified by the Glivenko-Cantelli Theorem, these tests use various measures of the "difference" between the EDF for the sample in question, and the cumulative distribution function for the hypothesized population. Well-known examples of EDF goodness-of-fit tests include the Kolmogorov-Smirnov test (Kolmogorov, 1933; Smirnov, 1948), the Cramér-von Mises test (Cramér, 1928; von Mises, 1928), and the test of Anderson and Darling (1954). Watson (1961) and Kuiper (1962) introduced related tests that are invariant to cyclic transformations of the data, thus broadening the range of situations in which they are applicable.

These EDF tests are designed for use with continuous data (as is Urzúa's LM test). Unfortunately, this fact is often overlooked, and researchers apply them (*inappropriately*) to discrete distributions. In this respect, the use of the Kolmogorov-Smirnov test by Clauset *et al*. (2009), and subsequent authors such as Brzezinski (2015), is questionable. The complication is that although test statistics based on the EDF are themselves distribution-free in the continuous case, this is not the case in general for discrete data (Conover, 1972). In such cases, EDF-based tests must be modified in order to be valid, but this point has received relatively limited attention in the literature. However, see Walsh (1963) and Henze (1996), Klar (1999), and Khmaladze (2013), for example.

Freedman (1981) proposes a modified form of Watson's (1961) U_N^2 statistic for use with discrete data. He also provides Monte Carlo evidence that his test out-performs Kuiper's (1962) modified test for various discrete distributions. We examine Freedman's test in the context of the Zipf distribution.

Let us simplify the notation by defining $p_i = Pr.[X = i] = (i^{-s}/H_{n,s})$; $i = 1, 2, 3, \ldots, n$; $s > 0$. We will be concerned with testing the null hypothesis, H_0: "The data follow a discrete distribution, *F*, defined by the probabilities $\{p_i\}_{i=1}^n$", against the alternative hypothesis, H_1: "H_0 is not true". If we have a sample of *N* independent observations, and $\{r_i\}_{i=1}^n$ are the sample frequencies, then $\sum_{i=1}^n r_i = N$, and Freedman's test statistic is:

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

where

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

Freedman shows that the asymptotic null distribution of the statistic in (2) is the same as for a weighted sum of (*n* - 1) independent chi-squared variates, each with one degree of freedom, and with weights which are the eigenvalues of a matrix with $(i, j)^{th}$ element given by

$$(p_i/n^2)\left(\{n - max(i,j)\}\, min(i,j) - \sum_{k=1}^{n-1} p_k \{n - max(i,j)\}\, min(j,k)\right) . \qquad (3)$$

Based on this result, Freedman *approximates* the quantiles of the asymptotic null distribution of U_N^{*2} by fitting Pearson curves. However, more recently Giles (2013) has shown that the complete asymptotic null distribution of this statistic can be obtained *exactly* by using standard computational methods. These techniques are those suggested by Imhof (1961), Davies (1973, 1980) and others, to invert the characteristic function for a statistic which is a weighted sum of chi-squared variates. *No approximations are needed*, and the numerical precision can be controlled quite simply.

Giles (2013) provides tables of quantiles for the null distribution of U_N^{*2} for several discrete distributions – namely, uniform, Benford, and beta-binomial. He also demonstrates that the test has excellent power properties against a range of alternatives. In this paper we extend his results to the case where the data are hypothesized to follow the Zipf distribution.

3. Results

3.1 Quantiles

The exact asymptotic distribution of U_N^{*2}, when the data are Zipf-distributed, is plotted in Figure 1, for various values of the scaling parameter, *s*, and for *n* = 10, 25. Table 1 reports the 90%, 95%, and 99% quantiles for the asymptotic null distribution of U_N^{*2}. The values of the scaling parameter, *s*, that are considered cover the range commonly encountered in empirical studies involving the Zipf distribution. (*e.g.*, see Clauset *et al*. 2009, p.662). A computed test statistic that exceeds the appropriate quantile value implies rejection of the hypothesis that the data follow a Zipf distribution.

These exact quantiles for the asymptotic distribution were computed by using Davies' (1980) method, *via* the psum.chisq command in the mgcv package in R (Wood, 2022), with the settings "tol = 10^{-6}" and "plim = 100,000". A simple modification of the code facilitates the computation of p-values associated with the U_N^{*2} test for the Zipf distribution. The R code that was used to produce the results in Table 1, as well as that for computing p-values, and the code for the simulation experiment discussed in the nest sub-section, can be downloaded from https://github.com/DaveGiles1949/r-code. It should be noted that if the value of *s* has to be estimated to apply the test, then one might use a log (rank-1/2) *vs*. log(size) OLS regression, as suggested by Gabaix and Ibragimov (2011), to reduce bias.

3.2 Power of the test

Figure 2 illustrates the exact asymptotic distribution of U_N^{*2} when the data follow other distributions. In Figure 2(a) the data follow a zero-truncated Poisson (ZTP) distribution, with n = 10, for various values of the distribution's parameter, λ. In Figure 2(b) the data follow the distribution for Benford's law for first digits, so *n* = 9. The power of the test based on U_N^{*2} has been investigated against these two alternative distributions, because their p.m.f's are visually very similar to that for the Zipf distribution for the given choices of *n* and λ. Moreover, the leading digits of data that satisfy Zipf's law with *s* = 1 satisfy Benford's law (Pietronero *et al*., 2001).

A small Monte Carlo simulation experiment has been conducted to examine the power properties of the new test against the alternative hypotheses of a ZTP distribution (with λ = 1), and Benford's Law first digits. The power of the test based on U_N^{*2} was examined for a range of values of the shape parameter (*s*) for the Zipf distribution; and for various sample sizes, *N*. Using 10,000 replications, the test was applied with the appropriate (asymptotic) quantiles from Table 1, corresponding to nominal significance levels of α = 10%, 5% and 1%. The "size distortion" was found to be negligible, even with samples as small as *N* = 10, as is illustrated in Table 2 for the case where *n* = 10.

The power results appear in Figures 3 and 4 for the ZTP and Benford's Law alternatives respectively. It should be noted that the curves shown in these two figures are *not* conventional power curves. The null and alternative hypotheses are non-nested. The horizontal axis measures the sample size, in each case, and *not* the departure from the null. These figures are illustrative and representative of a wider set of results that was generated. The test under consideration, and the quantiles used, are based on their asymptotic validity. However, the results show that the test has excellent power – in some cases with samples of only 20 to 40 observations. Although these results are specific to the chosen alternative distributions, they are most encouraging.

4. Empirical applications

In this section we illustrate the application of the proposed test to various actual data-sets, and compare the results with those that would be obtained if a researcher (wrongly) applied a conventional goodness-of-fit test based on the EDF for continuous data. The data used in these applications can be downloaded from https://github.com/DaveGiles1949/Data.

4.1 Terrorism events

The first application uses data for the number of deaths arising from terrorist attacks, as discussed by Clauset *et al*., (2007). The full sample of data comprises 13,274 observations, and the number of deaths per attack ranges from 1 to 2,749, with a median of 1 and a mean of 4.2. From (1), the log-likelihood function for *s*, based on a sample of *N* independent observations, is

$$l(s|\boldsymbol{x}) = \log\big(L(s|\boldsymbol{x})\big) = -s\sum_{i=1}^{N}\log(x_i) - Nlog\big(H_{n,s}\big) \quad . \tag{4}$$

Using this result. the maximum likelihood estimate, $\tilde{s}$, and the corresponding asymptotic standard error (a.s.e) are obtained using the 'mle2' function with the 'BFGS option', in the R package, 'sads' (Prado *et al.*, 2024).

Table 3 (a) reports the U_N^{*2} test results for the full sample of data and various sub-samples, the latter each beginning with the smallest ranked value in the full sample. The number of categories (*n*) increases with the sample size (*N*). This effectively provides various lower bounds for testing the validity of the Zipf distribution. Also reported in that table are values for the Kolmogorov-Smirnov (K-S) statistic for testing the Zipf hypothesis, and the associated p-values. These are provided to illustrate the consequences of applying this EDF test while ignoring the discrete nature of the data.

We follow the procedure suggested by Clauset *et al*. (2009, p.677), with 1,000 replications, to compute an approximate p-value for the K-S statistic. The U_N^{*2} test results in Table 3(a) reject the Zipf hypothesis in all cases, except when $n = 5$. In contrast, the K-S test fails to reject the Zipf hypothesis for every value of 'n'. This illustrates the consequences of using the latter test with discrete data. Figures 5 to 10 show the empirical histograms for the fatalities data for the different choices of 'n', together with plots of the logarithm of the data frequencies against the logarithms of their ranks. Informally, a negatively sloped straight line log-log plot suggests support for the Zipf distribution, and we see this most clearly in Figure 10(b).

4.2 Sports records

Our second application involves data for three sporting events – international cricket, the summer Olympic Games, and tennis Grand Slam tournaments. The first of these relates to centuries (scores of at least 100 "runs") scored by male cricketers in international "test" matches, so $n = 100$. See Wikipedia 1024a). The sample size is $N = 127$. The sample values range from 15 to 100 (centuries), with a median of 23 and a mean of 27.98. The second sports variable measures the total number of medals won by each of $N = 93$ countries at the 2020 summer Olympics. See Wikipedia (2024b). The number of medals won ranges from 1 to $n = 113$, with a median of 4, and a mean of 11.61. For the tennis data we have $N = 38$ males who achieved Grand Slam championship status at least four times, with a maximum of $n = 24$ wins in Grand Slam championship matches. So, the minimum number of wins is 4, with a median of 7 and mean of 7.95.

The maximum likelihood estimates of 's' and the asymptotic standard errors; together with the corresponding U_N^{*2} statistics and their p-values, are shown in Table 3(b). Again, the corresponding K-S test values and their approximate p-values are also reported there. Both the U_N^{*2} and K-S tests lead to a rejection of the Zipf hypothesis for the cricket data and the tennis data. In contrast, while the former test also suggests a rejection of this hypothesis in the case of the Olympics data, the K-S test's p-value implies a rejection at the 5% significance level, but not at the 1% level. Figures 11 to 13 provide the corresponding data histograms and the log(rank) - log(frequency) plots. While the latter may provide limited visual support for the Zipf distribution, this is offset by the formal goodness-of-fit test results.

5. Conclusions

The Zipf distribution, and "Zipf's Law" have been discussed extensively and applied to numerous data-sets across a broad range of disciplines. Given that this distribution is associated with discrete (rather than

continuous) data, particular care must be taken when constructing a formal test of the hypothesis that the data are Zipf-distributed. This point has essentially been ignored in the associated literature.

In this paper we propose the use of Freedman's (1981) modification of Watson's (1961) EDF test, together with the exact computations for the null distribution of the test statistic suggested by Giles (2013), in order to fill this gap in the literature. The exact quantiles of the asymptotic null distribution of the new test statistic are reported for various values of the Zipf distribution's scaling parameter, and a range of values for the number of categories associated with the data. These quantiles facilitate the application of test by other researchers.

Although the test is, strictly, one with asymptotic justification, we provide simulation-based evidence that verifies that it has excellent power against various alternative hypotheses, even with relatively small sample sizes. This adds further support for the use of the test that has been proposed.

Several applications with various data-sets have been to illustrate how the test performs in practice. The results of these applications also demonstrate that the mis-use of a familiar EDF test that is valid only for *continuous* data can lead to conflicting conclusions.

Table 1: Quantiles for the asymptotic null distribution of U_N^{*2} *

	s = 0.50			*s = 0.75*		
n	***0.90***	***0.95***	***0.99***	***0.90***	***0.95***	***0.99***
2	0.1641	0.2330	0.4025	0.1582	0.2246	0.3879
3	0.1663	0.2172	0.3370	0.1608	0.2111	0.3307
4	0.1614	0.2052	0.3073	0.1566	0.2002	0.3033
5	0.1577	0.1981	0.2925	0.1531	0.1934	0.2886
6	0.1552	0.1938	0.2840	0.1505	0.1890	0.2798
7	0.1534	0.1910	0.2788	0.1486	0.1860	0.2740
8	0.1522	0.1891	0.2753	0.1472	0.1837	0.2700
9	0.1512	0.1877	0.2728	0.1460	0.1820	0.2670
10	0.1505	0.1866	0.2710	0.1451	0.1807	0.2647
11	0.1500	0.1858	0.2696	0.1443	0.1796	0.2629
12	0.1495	0.1852	0.2686	0.1436	0.1786	0.2614
13	0.1492	0.1846	0.2677	0.1430	0.1779	0.2601
14	0.1488	0.1842	0.2670	0.1425	0.1772	0.2591
15	0.1486	0.1838	0.2664	0.1421	0.1766	0.2582
16	0.1483	0.1835	0.2659	0.1416	0.1760	0.2573
17	0.1481	0.1832	0.2655	0.1413	0.1755	0.2566
18	0.1480	0.1830	0.2651	0.1409	0.1751	0.2560
19	0.1478	0.1828	0.2648	0.1406	0.1747	0.2554
20	0.1477	0.1826	0.2645	0.1403	0.1743	0.2549
30	0.1467	0.1814	0.2628	0.1382	0.1717	0.2513
40	0.1462	0.1808	0.2620	0.1368	0.1701	0.2492
50	0.1459	0.1804	0.2615	0.1358	0.1689	0.2477
75	0.1454	0.1799	0.2609	0.1341	0.1669	0.2453
100	0.1451	0.1795	0.2605	0.1330	0.1656	0.2437
200	0.1445	0.1789	0.2599	0.1305	0.1629	0.2404
300	0.1442	0.1786	0.2596	0.1292	0.1614	0.2387

* H_0: "The data follow a Zipf distribution"; *n* = maximum value of *X*; *s* = scaling parameter

Table 1 (continued): Quantiles for the asymptotic null distribution of U_N^{*2} *

	s = 1.00			*s = 1.25*		
n	***0.90***	***0.95***	***0.99***	***0.90***	***0.95***	***0.99***
2	0.1503	0.2134	0.3686	0.1409	0.2001	0.3456
3	0.1534	0.2024	0.3207	0.1442	0.1915	0.3070
4	0.1497	0.1927	0.2958	0.1408	0.1828	0.2845
5	0.1463	0.1862	0.2814	0.1374	0.1763	0.2703
6	0.1436	0.1816	0.2721	0.1343	0.1713	0.2603
7	0.1414	0.1781	0.2656	0.1317	0.1673	0.2529
8	0.1396	0.1754	0.2608	0.1294	0.1639	0.2471
9	0.1380	0.1732	0.2570	0.1274	0.1611	0.2422
10	0.1367	0.1713	0.2539	0.1256	0.1586	0.2381
11	0.1356	0.1697	0.2514	0.1240	0.1564	0.2346
12	0.1345	0.1684	0.2492	0.1225	0.1544	0.2314
13	0.1336	0.1671	0.2473	0.1211	0.1526	0.2286
14	0.1327	0.1660	0.2456	0.1198	0.1509	0.2261
15	0.1320	0.1650	0.2441	0.1186	0.1494	0.2237
16	0.1312	0.1641	0.2427	0.1175	0.1480	0.2216
17	0.1306	0.1632	0.2414	0.1165	0.1466	0.2196
18	0.1299	0.1624	0.2403	0.1155	0.1454	0.2177
19	0.1293	0.1617	0.2392	0.1146	0.1442	0.2159
20	0.1288	0.1610	0.2382	0.1137	0.1431	0.2196
30	0.1245	0.1557	0.2308	0.1066	0.1343	0.2079
40	0.1215	0.1521	0.2259	0.1016	0.1281	0.1926
50	0.1193	0.1494	0.2222	0.0978	0.1234	0.1857
75	0.1152	0.1446	0.2157	0.0909	0.1149	0.1734
100	0.1124	0.1412	0.2112	0.0861	0.1089	0.1649
200	0.1060	0.1335	0.2006	0.0751	0.0953	0.1449
300	0.1024	0.1292	0.1947	0.0690	0.0877	0.1338

* H_0: "The data follow a Zipf distribution"; n = maximum value of X; s = scaling parameter

Table 1 (continued): Quantiles for the asymptotic null distribution of U_N^{*2} *

	s = 1.50			*s = 1.75*		
n	***0.90***	***0.95***	***0.99***	***0.90***	***0.95***	***0.99***
2	0.1305	0.1853	0.3201	0.1195	0.1696	0.2930
3	0.1337	0.1788	0.2896	0.1223	0.1647	0.2693
4	0.1304	0.1707	0.2692	0.1189	0.1570	0.2506
5	0.1267	0.1641	0.2551	0.1149	0.1501	0.2364
6	0.1233	0.1586	0.2445	0.1110	0.1441	0.2251
7	0.1202	0.1539	0.2360	0.1074	0.1388	0.2157
8	0.1173	0.1499	0.2290	0.1041	0.1342	0.2078
9	0.1148	0.1463	0.2231	0.1011	0.1300	0.2008
10	0.1125	0.1432	0.2179	0.0983	0.1262	0.1946
11	0.1103	0.1403	0.2132	0.0958	0.1228	0.1890
12	0.1084	0.1377	0.2091	0.0934	0.1197	0.1840
13	0.1065	0.1353	0.2053	0.0913	0.1168	0.1794
14	0.1048	0.1331	0.2018	0.0892	0.1141	0.1752
15	0.1032	0.1310	0.1986	0.0873	0.1116	0.1712
16	0.1017	0.1290	0.1956	0.0855	0.1093	0.1676
17	0.1003	0.1272	0.1928	0.0839	0.1071	0.1642
18	0.0990	0.1255	0.1902	0.0823	0.1051	0.1610
19	0.0977	0.1239	0.1877	0.0808	0.1031	0.1580
20	0.0965	0.1223	0.1854	0.0794	0.1013	0.1551
30	0.0870	0.1103	0.1671	0.0684	0.0871	0.1333
40	0.0803	0.1018	0.1545	0.0609	0.0775	0.1185
50	0.0752	0.0954	0.1449	0.0553	0.0705	0.1077
75	0.0663	0.0842	0.1281	0.0460	0.0585	0.0894
100	0.0603	0.0766	0.1167	0.0400	0.0509	0.0777
200	0.0472	0.0600	0.0917	0.0278	0.0354	0.0540
300	0.0405	0.0515	0.0788	0.0222	0.0282	0.0430

* H_0: "The data follow a Zipf distribution"; n = maximum value of X; s = scaling parameter

Table 1 (continued): Quantiles for the asymptotic null distribution of U_N^{*2} *

	s = 2.00			*s = 2.25*		
n	***0.90***	***0.95***	***0.99***	***0.90***	***0.95***	***0.99***
2	0.1082	0.1537	0.2654	0.0971	0.1378	0.2381
3	0.1106	0.1498	0.2469	0.0988	0.1347	0.2235
4	0.1070	0.1423	0.2296	0.0950	0.1273	0.2072
5	0.1026	0.1351	0.2154	0.0903	0.1199	0.1931
6	0.0983	0.1287	0.2036	0.0857	0.1132	0.1811
7	0.0943	0.1229	0.1935	0.0815	0.1072	0.1707
8	0.0906	0.1178	0.1848	0.0777	0.1019	0.1616
9	0.0873	0.1133	0.1771	0.0743	0.0971	0.1536
10	0.0843	0.1091	0.1703	0.0711	0.0929	0.1466
11	0.0815	0.1054	0.1641	0.0682	0.0890	0.1402
12	0.0789	0.1019	0.1585	0.0656	0.0855	0.1344
13	0.0765	0.0988	0.1534	0.0632	0.0822	0.1292
14	0.0743	0.0958	0.1488	0.0610	0.0793	0.1244
15	0.0723	0.0931	0.1444	0.0589	0.0766	0.1201
16	0.0704	0.0906	0.1404	0.0570	0.0741	0.1160
17	0.0686	0.0883	0.1367	0.0553	0.0717	0.1123
18	0.0669	0.0861	0.1332	0.0536	0.0695	0.1088
19	0.0653	0.0840	0.1300	0.0521	0.0675	0.1056
20	0.0638	0.0820	0.1269	0.0507	0.0656	0.1025
30	0.0525	0.0673	0.1038	0.0400	0.0516	0.0803
40	0.0451	0.0578	0.0890	0.0333	0.0429	0.0666
50	0.0398	0.0510	0.0784	0.0287	0.0369	0.0572
75	0.0313	0.0400	0.0614	0.0215	0.0276	0.0427
100	0.0261	0.0333	0.0511	0.0174	0.0223	0.0343
200	0.0164	0.0208	0.0318	0.0101	0.0128	0.0196
300	0.0123	0.0156	0.0237	0.0072	0.0091	0.0139

* H_0: "The data follow a Zipf distribution"; n = maximum value of X; s = scaling parameter

Table 1 (continued): Quantiles for the asymptotic null distribution of U_N^{*2} *

	s = 2.50			*s = 2.75*		
n	***0.90***	***0.95***	***0.99***	***0.90***	***0.95***	***0.99***
2	0.0863	0.1226	0.2117	0.0762	0.1082	0.1869
3	0.0875	0.1199	0.2000	0.0768	0.1057	0.1771
4	0.0834	0.1125	0.1847	0.0726	0.0985	0.1627
5	0.0786	0.1051	0.1708	0.0677	0.0912	0.1494
6	0.0739	0.0984	0.1589	0.0631	0.0846	0.1379
7	0.0697	0.0924	0.1485	0.0590	0.0788	0.1279
8	0.0659	0.0871	0.1395	0.0553	0.0737	0.1193
9	0.0624	0.0824	0.1316	0.0521	0.0692	0.1117
10	0.0593	0.0781	0.1246	0.0492	0.0653	0.1051
11	0.0565	0.0744	0.1184	0.0466	0.0617	0.0993
12	0.0540	0.0709	0.1128	0.0442	0.0586	0.0940
13	0.0517	0.0679	0.1078	0.0421	0.0557	0.0894
14	0.0496	0.0650	0.1032	0.0402	0.0531	0.0851
15	0.0477	0.0625	0.0990	0.0385	0.0508	0.0813
16	0.0459	0.0601	0.0951	0.0369	0.0486	0.0778
17	0.0443	0.0579	0.0916	0.0354	0.0467	0.0746
18	0.0427	0.0559	0.0883	0.0340	0.0449	0.0717
19	0.0413	0.0540	0.0853	0.0328	0.0432	0.0690
20	0.0400	0.0523	0.0825	0.0316	0.0417	0.0665
30	0.0305	0.0397	0.0624	0.0234	0.0308	0.0489
40	0.0248	0.0322	0.0504	0.0187	0.0245	0.0388
50	0.0209	0.0271	0.0424	0.0155	0.0203	0.0322
75	0.0152	0.0196	0.0306	0.0110	0.0143	0.0226
100	0.0119	0.0154	0.0240	0.0085	0.0111	0.0175
200	0.0065	0.0084	0.0130	0.0045	0.0059	0.0092
300	0.0045	0.0058	0.0090	0.0031	0.0040	0.0063

* H_0: "The data follow a Zipf distribution"; n = maximum value of X; s = scaling parameter

Table 1 (continued): Quantiles for the asymptotic null distribution of U_N^{*2} *

	s = 3.00			*s = 3.25*		
n	***0.90***	***0.95***	***0.99***	***0.90***	***0.95***	***0.99***
2	0.0668	0.0949	0.1638	0.0582	0.0827	0.1428
3	0.0669	0.0925	0.1555	0.0579	0.0803	0.1354
4	0.0626	0.0855	0.1420	0.0537	0.0737	0.1229
5	0.0579	0.0785	0.1294	0.0492	0.0671	0.1113
6	0.0535	0.0722	0.1185	0.0452	0.0613	0.1012
7	0.0496	0.0668	0.1092	0.0416	0.0563	0.0927
8	0.0462	0.0620	0.1012	0.0385	0.0520	0.0854
9	0.0432	0.0579	0.0942	0.0358	0.0483	0.0791
10	0.0406	0.0543	0.0882	0.0334	0.0450	0.0737
11	0.0382	0.0511	0.0828	0.0313	0.0422	0.0690
12	0.0361	0.0482	0.0781	0.0295	0.0396	0.0648
13	0.0342	0.0457	0.0739	0.0279	0.0374	0.0611
14	0.0325	0.0434	0.0701	0.0264	0.0354	0.0578
15	0.0310	0.0413	0.0667	0.0251	0.0336	0.0548
16	0.0296	0.0394	0.0636	0.0239	0.0320	0.0521
17	0.0283	0.0377	0.0608	0.0228	0.0305	0.0497
18	0.0272	0.0361	0.0583	0.0218	0.0292	0.0475
19	0.0261	0.0347	0.0559	0.0209	0.0280	0.0455
20	0.0251	0.0333	0.0537	0.0200	0.0268	0.0436
30	0.0182	0.0241	0.0387	0.0143	0.0191	0.0310
40	0.0143	0.0189	0.0303	0.0111	0.0148	0.0240
50	0.0118	0.0156	0.0249	0.0091	0.0121	0.0196
75	0.0082	0.0108	0.0173	0.0063	0.0083	0.0135
100	0.0063	0.0083	0.0132	0.0048	0.0064	0.0103
200	0.0033	0.0043	0.0068	0.0025	0.0033	0.0053
300	0.0022	0.0029	0.0046	0.0016	0.0022	0.0035

* H_0: "The data follow a Zipf distribution"; n = maximum value of X; s = scaling parameter

Table 1 (continued): Quantiles for the asymptotic null distribution of U_N^{*2} *

	s = 3.50			*s = 3.75*		
n	***0.90***	***0.95***	***0.99***	***0.90***	***0.95***	***0.99***
2	0.0505	0.0717	0.1238	0.0436	0.0618	0.1068
3	0.0499	0.0693	0.1172	0.0427	0.0595	0.1009
4	0.0458	0.0631	0.1057	0.0390	0.0538	0.0905
5	0.0417	0.0571	0.0951	0.0351	0.0483	0.0809
6	0.0379	0.0518	0.0860	0.0318	0.0436	0.0728
7	0.0347	0.0473	0.0783	0.0289	0.0396	0.0660
8	0.0320	0.0434	0.0718	0.0265	0.0362	0.0602
9	0.0296	0.0401	0.0663	0.0244	0.0334	0.0554
10	0.0275	0.0373	0.0615	0.0227	0.0309	0.0512
11	0.0257	0.0348	0.0573	0.0211	0.0288	0.0476
12	0.0241	0.0326	0.0537	0.0197	0.0269	0.0445
13	0.0227	0.0307	0.0505	0.0185	0.0252	0.0418
14	0.0214	0.0290	0.0476	0.0175	0.0238	0.0393
15	0.0203	0.0275	0.0451	0.0165	0.0225	0.0372
16	0.0193	0.0261	0.0428	0.0157	0.0213	0.0352
17	0.0184	0.0248	0.0407	0.0149	0.0203	0.0335
18	0.0175	0.0237	0.0389	0.0142	0.0193	0.0319
19	0.0168	0.0226	0.0371	0.0136	0.0184	0.0304
20	0.0161	0.0217	0.0356	0.0130	0.0176	0.0291
30	0.0113	0.0153	0.0250	0.0091	0.0123	0.0203
40	0.0088	0.0118	0.0193	0.0070	0.0095	0.0156
50	0.0071	0.0096	0.0157	0.0057	0.0077	0.0126
75	0.0049	0.0066	0.0107	0.0039	0.0052	0.0086
100	0.0037	0.0050	0.0081	0.0029	0.0040	0.0065
200	0.0019	0.0025	0.0041	0.0015	0.0020	0.0033
300	0.0013	0.0017	0.0028	0.0010	0.0013	0.0022

* H_0: "The data follow a Zipf distribution"; n = maximum value of X; s = scaling parameter

Table 2: True sizes of the U_N^{*2} test (n = 10)*

***N*:**	10	20	30	10	20	30
α		***s* = 0.5**			***s* = 0.75**	
0.10	0.095	0.099	0.102	0.095	0.094	0.103
0.05	0.046	0.050	0.051	0.048	0.051	0.048
0.01	0.008	0.009	0.009	0.008	0.009	0.008
α		***s* = 1.0**			***s* = 1.25**	
0.10	0.098	0.099	0.096	0.101	0.100	0.094
0.05	0.047	0.050	0.047	0.048	0.051	0.046
0.01	0.008	0.009	0.009	0.007	0.011	0.010
α		***s* = 1.5**			***s* = 1.75**	
0.10	0.099	0.096	0.093	0.101	0.098	0.095
0.05	0.049	0.050	0.046	0.047	0.051	0.047
0.01	0.009	0.012	0.009	0.010	0.011	0.009
α		***s* = 2.0**			***s* = 2.25**	
0.10	0.094	0.097	0.095	0.103	0.095	0.095
0.05	0.048	0.051	0.050	0.049	0.051	0.049
0.01	0.010	0.012	0.010	0.012	0.010	0.010
α		***s* = 2.5**			***s* = 2.75**	
0.10	0.094	0.095	0.093	0.120	0.101	0.092
0.05	0.046	0.050	0.046	0.045	0.049	0.049
0.01	0.013	0.013	0.010	0.014	0.013	0.010
α		***s* = 3.0**			***s* = 3.25**	
0.10	0.089	0.084	0.085	0.080	0.102	0.101
0.05	0.044	0.054	0.051	0.054	0.051	0.049
0.01	0.014	0.016	0.010	0.015	0.016	0.012
α		***s* = 3.5**			***s* = 3.75**	
0.10	0.089	0.072	0.089	0.065	0.085	0.105
0.05	0.043	0.056	0.058	0.051	0.040	0.040
0.01	0.022	0.019	0.012	0.018	0.018	0.012

***** α = Nominal significance level of the test

Table 3: Test of Zipf distribution*

(a) Terrorist fatalities data

n	N	$\tilde{s}$ (a.s.e)	U_N^{*2}	p-val.	K-S	p-val.
2,749	13,274	1.8831 (0.0082)	0.0975	0.0000	0.0446	0.4999
400	13,272	1.8688 (0.0085)	0.3655	0.0000	0.0406	0.4831
50	13,171	1.8153 (0.0096)	0.7422	0.0000	0.0278	0.5046
20	12,926	1.7606 (0.0110)	0.5156	0.0000	0.0171	0.4216
10	12,420	1.7137 (0.0132)	0.3076	0.0009	0.0113	0.3759
5	11,461	1.6547 (0.0176)	0.0583	0.3483	0.0069	0.4507
		(b) Sports data				
Cricket						
100	127	0.3637 (0.0750)	3.2319	0.0000	0.0814	0.0000
Olympics						
113	95	1.2037 (0.0716)	1.0050	0.0000	0.0928	0.0336
Tennis						
24	38	0.4230 (0.1677)	0.8128	0.0000	0.1799	0.0000

* "a.s.e" denotes "asymptotic standard error

Figure 1a: Asymptotic Distribution of Test Statistic (Zipf Data: n = 10)

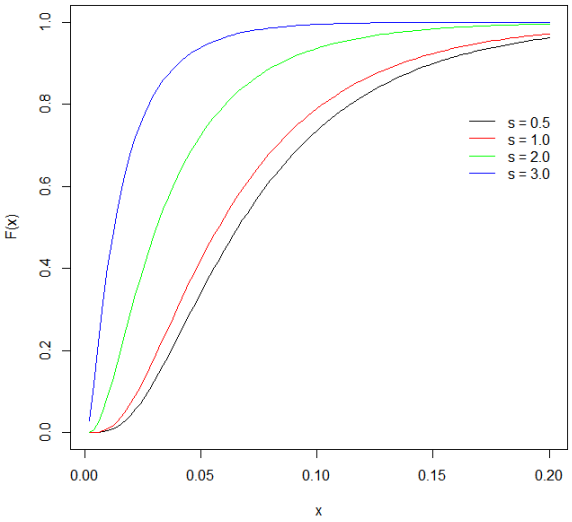


Figure 1b: Asymptotic Distribution of Test Statistic (Zipf Data: n = 25)

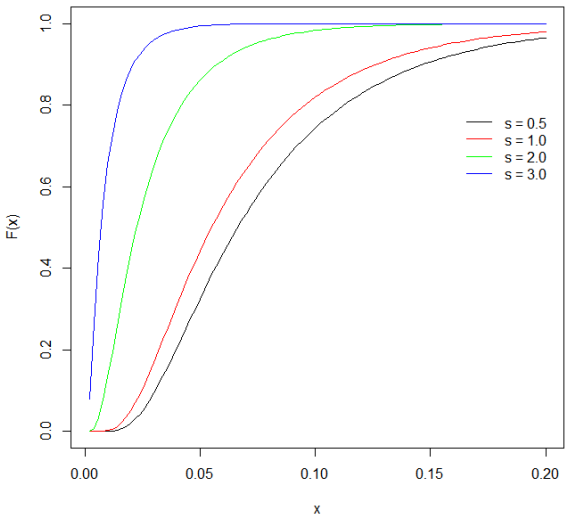


Figure 2a: Asymptotic Distribution of Test Statistic (ZTP Data: n = 10)

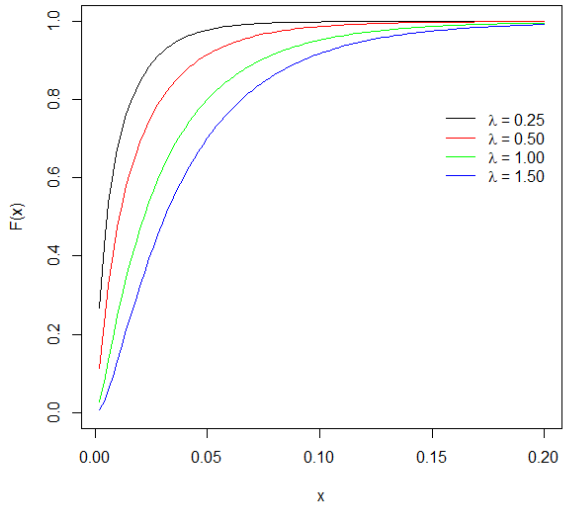


Figure 2b: Asymptotic Distribution of Test Statistic (Benford Data: n = 9)

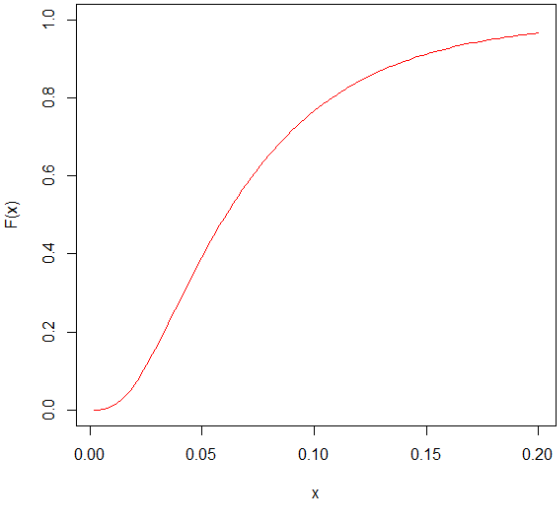


Figure 3(a): Powers of Test
(ZTP Alternative; s = 1.0)

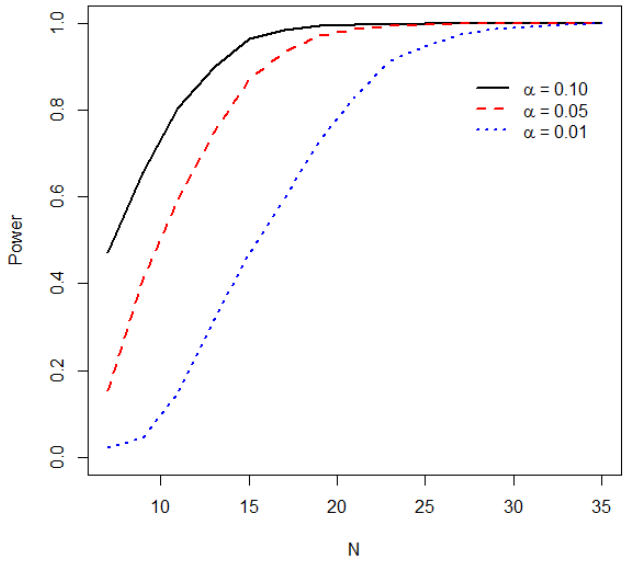


Figure 3(b): Powers of Test
(ZTP Alternative; s = 1.5)

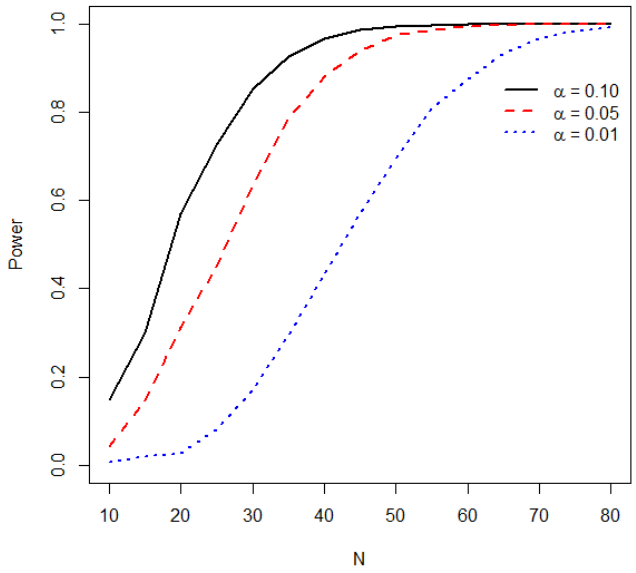


Figure 3(c): Powers of Test
(ZTP Alternative; s = 2.0)

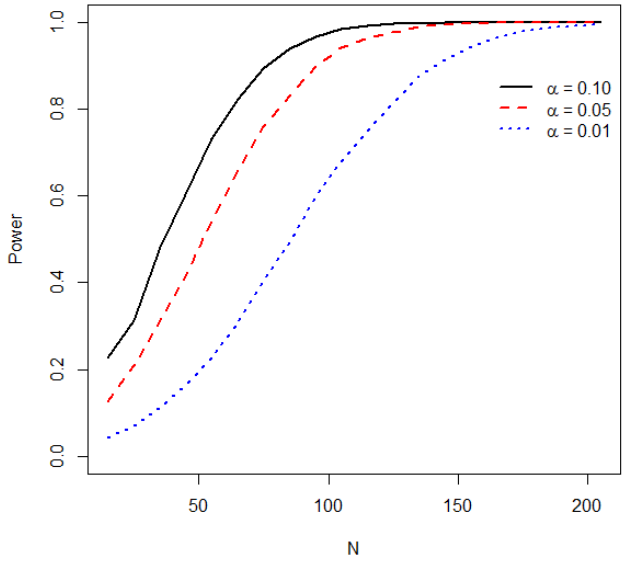


Figure 3(d): Powers of Test
(ZTP Alternative; s = 2.5)

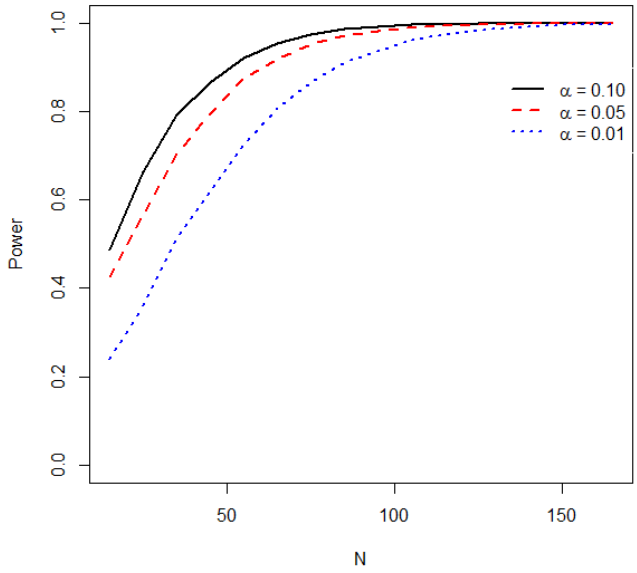


Figure 4(a): Powers of Test
(Benford Alternative; s = 1.0)

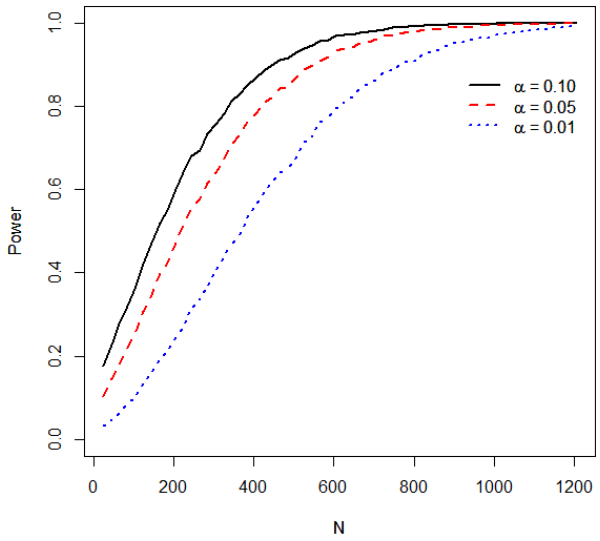


Figure 4(b): Powers of Test
(Benford Alternative; s = 1.25)

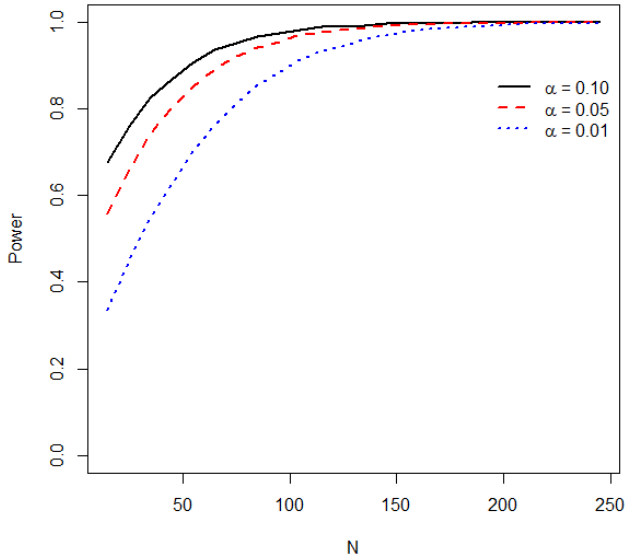


Figure 4(c): Powers of Test
(Benford Alternative; s = 1.5)

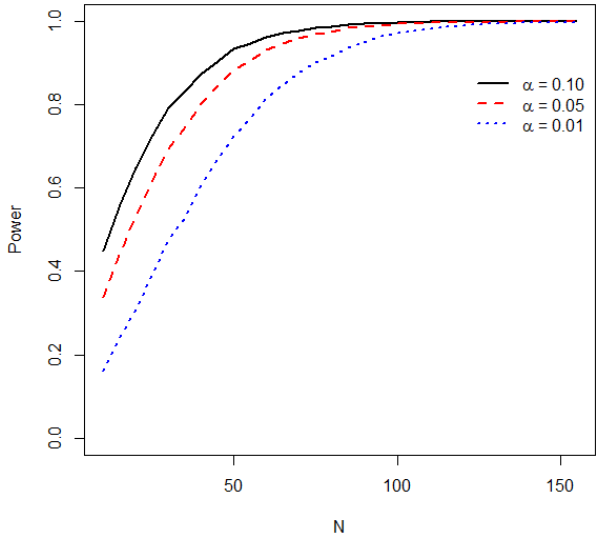


Figure 4(d): Powers of Test
(Benford Alternative; s = 2.0)

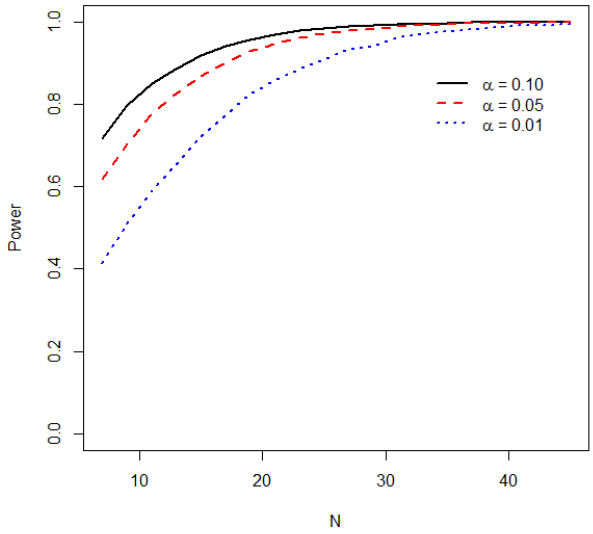


Figure 5(a): Terrorism Deaths (n = 2,749)

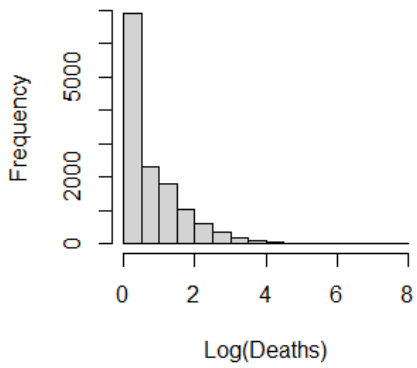


Figure 5(b): Terrorism Deaths (n = 2,749)

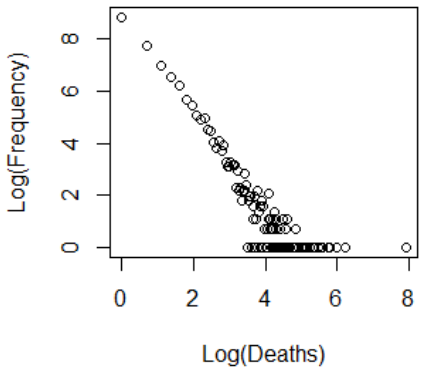


Figure 6(a): Terrorism Deaths (n = 400)

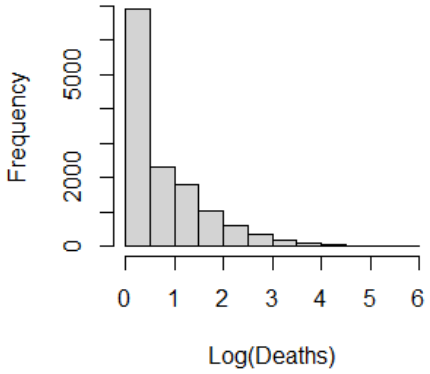


Figure 6(b): Terrorism Deaths (n = 400)

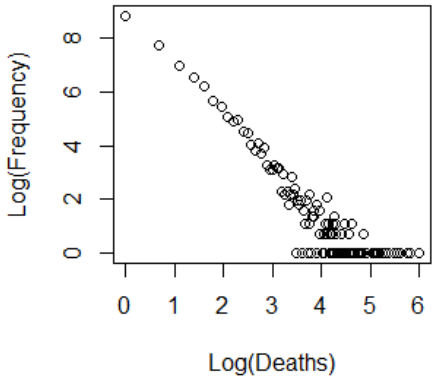


Figure 7(a): Terrorism Deaths (n = 50)

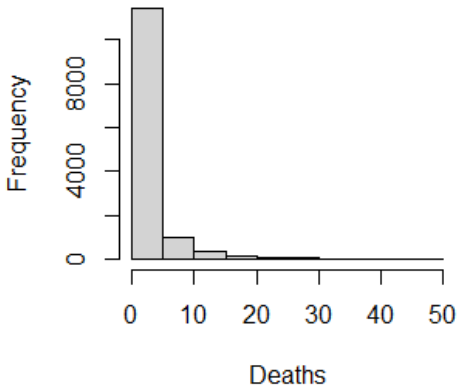


Figure 7(b): Terrorism Deaths (n = 50)

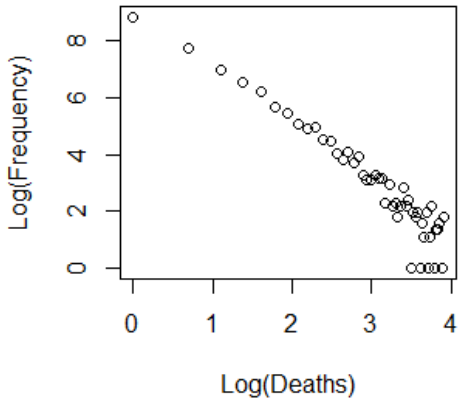


Figure 8(a): Terrorism Deaths (n = 20)

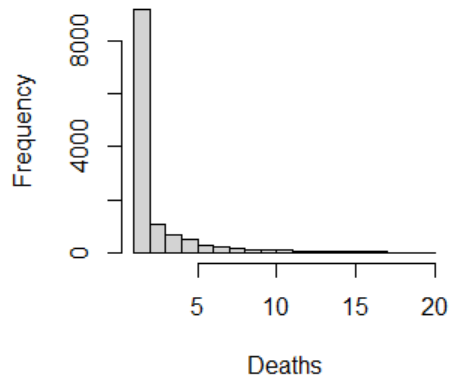


Figure 8(b): Terrorism Deaths (n = 20)

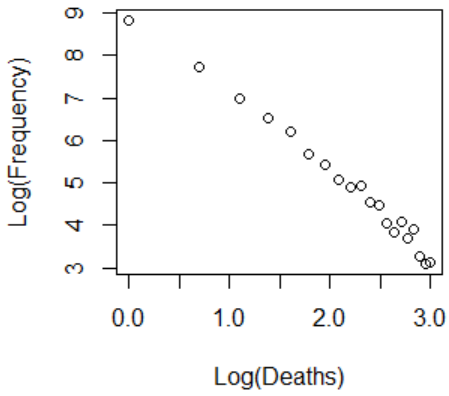


Figure 9(a): Terrorism Deaths (n = 10)

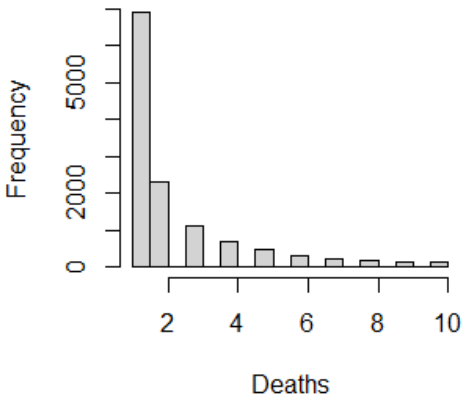


Figure 9(b): Terrorism Deaths (n = 10)

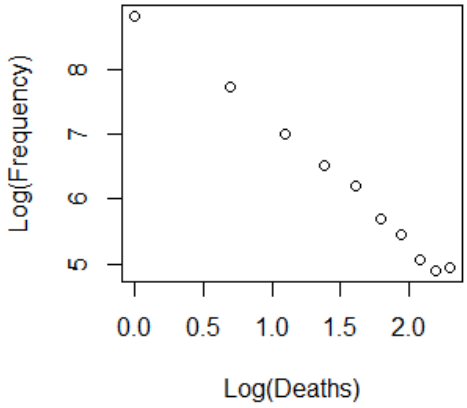


Figure 10(a): Terrorism Deaths (n = 5)

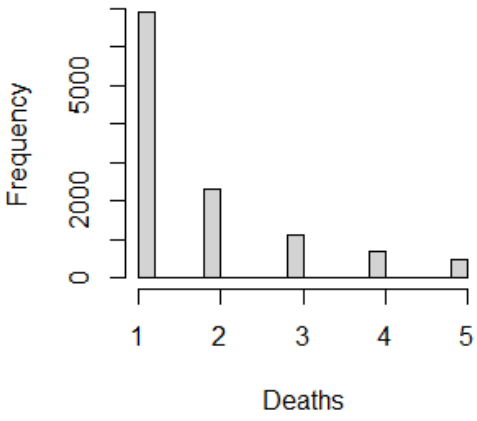


Figure 10(b): Terrorism Deaths (n = 5)

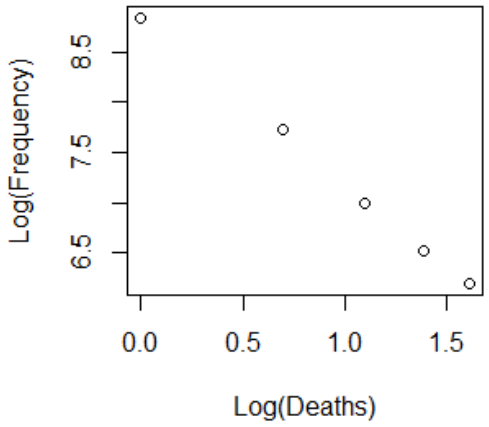


Figure 11(a): Cricket Centuries (Men)

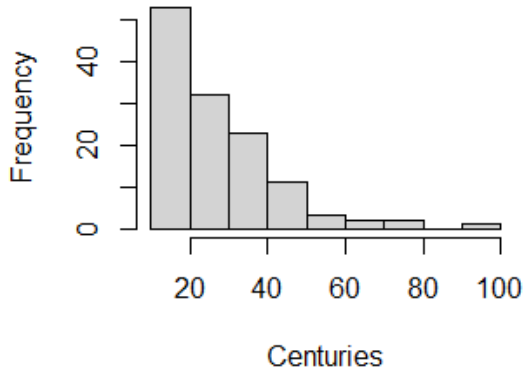


Figure 11(b): Cricket Centuries (Men)

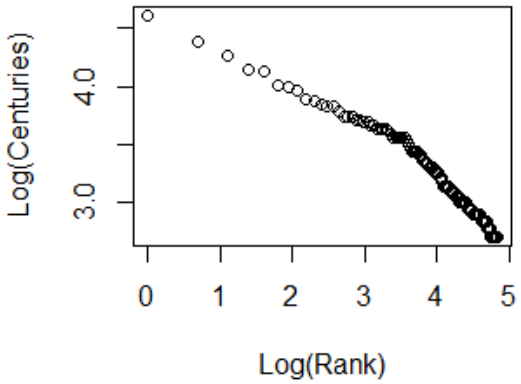


Figure 12(a): Olympic Medals (2020)

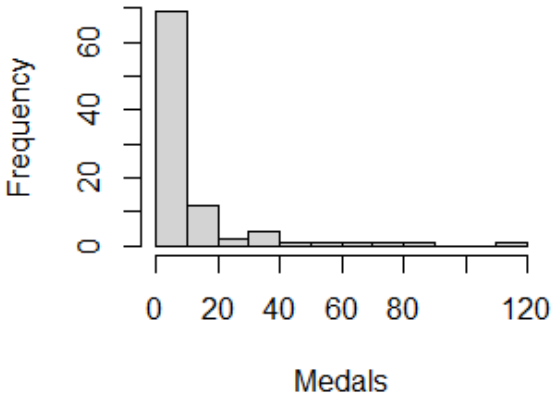


Figure 12(b): Olympic Medals (2020)

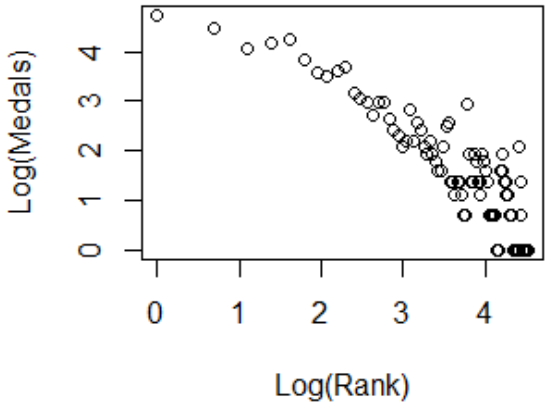


Figure 13(a): Tennis Grand Slams (Men)

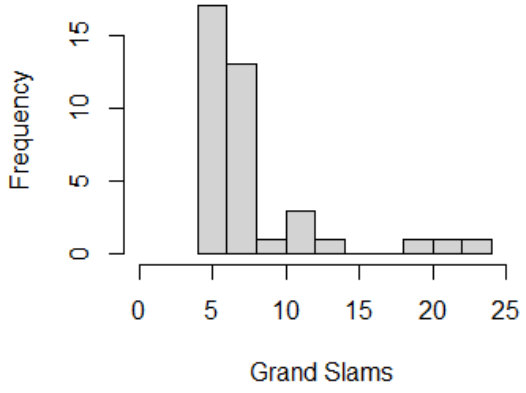
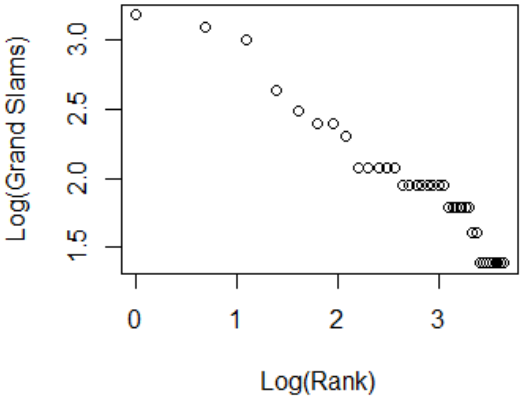


Figure 13(b): Tennis Grand Slams (Men)



References

Anderson, T. W. and D. A. Darling, 1954. A test of goodness-of-fit". *Journal of the American Statistical Association*, 49, 765–769.

Boamah-Addo, K, T. J. Kozubowski, and A. K. Panorska, 2022. A discrete truncated Zipf distribution. *Statistica Neerlandica*, 77, 156-187.

Brzezinski, M., 2015. Power laws in citation distributions: evidence from Scopus. *Scientometrics*, 103, 213-228.

Clauset, A., M. Young, and K. S. Gleditsch, 2007. On the frequency of severe terrorist events. *Journal of Conflict Resolution*, 51, 58-87.

Clauset, A., C. R. Shalizi, and M. E. J. Newman, 2009. Power-law distributions in empirical data. *SIAM Review*, 51. 661-703.

Conover, W. J., 1972. A Kolmogorov goodness-of-fit test for discontinuous distributions. *Journal of the American Statistical Association*, 67, 591-596.

Cramér, H., 1928. On the composition of elementary errors. *Scandinavian Actuarial Journal*, 1928, 13–74.

D'Agostino, R. B. and M. A. Stephens (eds.), 1986. *Goodness-of-Fit Techniques*. Marcel Dekker, New York.

Davies, R. B., 1973. Numerical inversion of a characteristic function. *Biometrika*, 60, 415-417.

Davies, R. B., 1980. The distribution of a linear combination of χ^2 random variables, algorithm AS 155. *Applied Statistics*, 29, 323-333.

Freedman, L. S., 1981. Watson's U_N^2 statistic for a discrete distribution. *Biometrika*, 68, 708-711.

Gabaix, X., 1999. Zipf 's law for cities: An explanation. *Quarterly Journal of Economics*, 114, 739–767.

Gabaix, X, 2009. Power laws in economics and finance. *Annual Review of Economics*, 1, 255–293.

Gabaix, X., 2016. Power laws in economics: An introduction. *Journal of Economic Perspectives*, 30, 185-206.

Gabaix, X. and R. Ibragimov, 2011. Rank-1/2: a simple way to improve the OLS estimation of tail exponents. *Journal of Business and Economic Statistics*, 29, 24–39.

Giesen, K. and J. Südekum, 2011. Zipf's law for cities in the regions and the country. *Journal of Economic Geography*, 11,667–686.

Giles, D. E., 2013, Exact asymptotic goodness-of-fit testing for discrete circular data, with applications. *Chilean Journal of Statistics*, 4, 19-34.

Henze, N., 1996. Empirical-distribution-function goodness-of-fit tests for discrete models. *Canadian Journal of Statistics*, 24, 81-93.

Hinloopen, J. and C. van Marrewijk, 2012. Power laws and comparative advantage. *Applied Economics*, 44, 1483-1507.

Imhof, J. P., 1961. Computing the distribution of quadratic forms in normal variables. *Biometrika*, 48, 419-426.

Khmaladze, E., 2013. A note on distribution free testing for discrete distributions, *Annals of Statistics*, 41, 2979-2993.

Klar, B., 1999. Goodness-of-fit tests for discrete models based on the integrated distribution function. *Metrika*, 49,53-69.

Kolmogorov, A., 1933. Sulla determinazione empirica di una legge di distribuzione. *Giornale dell'Istituto Italiano degli Attuari*, 4, 83–91.

Kuiper, N. H., 1962. Tests concerning random points on a circle. *Proceedings of the Koninklijke Nederlandse Akademie van Wetenschappen A*, 63, 38-47.

Mises, R. E. von, 1928. *Wahrscheinlichkeit, Statistik und Wahrheit*. Julius Springer, Vienna.

Naldi, N., 2015. Approximation of the truncated Zeta distribution and Zipf's law. arXiv:1511.01480v1 [stat.AP].

Newman, M. E. J., 2005. Power laws, Pareto distributions and Zip's law. *Contemporary Physics*, 46, 323-351.

Pietronero, L., E. Tosatti, V. Tosatti,and A. Vespignani, 2001. Explaining the uneven distribution of numbers in nature: The laws of Benford and Zipf. *Physica A*, 293, 297–304.

Prado, P. I., M. D. Miranda, and A. Chalom (2024). Package 'sads'. https://cran.r-project.org/web/packages/sads/sads.pdf.

Smirnov, N., 1948. Table for estimating the goodness of fit of empirical distributions. *Annals of Mathematical Statistics*, 19, 279–281.

Stephens, M. A., 1974. EDF statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association*, 69, 730–737.

Urzúa, C. M, 2000. A simple and efficient test for Zipf's law. *Economics Letters*, 66, 257-260.

Watson, G. S., 1961. Goodness-of-fit tests on a circle. I. *Biometrika*, 48, 109-114.

Wikipedia, 2024a. List of cricketers by number of international centuries scored. https://en.wikipedia.org/wiki/List_of_cricketers_by_number_of_international_centuries_scored

Wikipedia, 2024b. 2020 summer Olympics medal table. https://en.wikipedia.org/wiki/2020_Summer_Olympics_medal_table

Wikipedia, 2024c. List of Grand Slam men's single champions.

https://en.wikipedia.org/wiki/List_of_Grand_Slam_men%27s_singles_champions

Wood, S., 2022. Package 'mgcv': Mixed GAM computation vehicle with automatic smoothness estimation. https://cran.r-project.org/web/packages/mgcv/mgcv.pdf

Zipf, G. K., 1949. *Human Behavior and the Principle of Least Effort*. Addison-Wesley, Cambridge.