Aarhus, 20/9-2010

A check of Maddison's gdp data

Benford's Law with some range problems

Martin Paldam, School of Economics and Management, Aarhus University[1]

Abstract:

The paper studies if the Maddison set of data for GDP per capita follows a simple statistical regularity, known as Benford's Law. It is a simple logarithmic relation on the frequency of the first digit in a data set. These data ought to follow the law as they are Maddison's calibration of data compiled by many independent agencies and researches. The data set consists of 12,411 observations, so a rather strong test can be made. On visual inspection the relation appears to fit rather well, but the law is rejected by a formal test. The explanation of the rejection is found to be that the range of the data is too small.

Acknowledgement: I want to thank Tomasz Michalski for an inspiring presentation at the DEGIT 2010 conference at the Goethe University in Frankfurt.

1. School of Economics and Management, Bartholins Allé 10, Aarhus University, 8000 Aarhus C, Denmark. E-mail: mpaldam@econ.au.dk. URL: http://www.martin.paldam.dk

1. Introduction

The purpose of this note is to submit an important data set to a test. The data is Maddison's gdp set, which consists of 12,411 observations.[2] It is tested if these data follow a simple numerical regularity known as *Benford's Law*. It predicts that the *first digit* of these data should have the frequency distribution given in Table 1.

Table 1. Benford's distribution

Digit	Frequency	Digit	Frequency	Digit	Frequency
1	30.10 %	4	9.69 %	7	5.80 %
2	17.61 %	5	7.92 %	8	5.12 %
3	12.49 %	6	6.69 %	9	4.58 %

Three terms are used in a special way: (i) The lower case *gdp* means GDP per capita. The data are published as an integer in comparable 1990 US $.[3] (ii) The first digit of the gdp considers numbers in three *log-decades:* The log-decade of the 100s, where the first digit is followed by two more; the log decade of the 1000s, where the first digit is followed by three more; and finally the log-decade of the 10000s, where the first digit is followed by four more. (iii) The *range problem* occurs if the sample does not contain enough data in each log-decade for the distribution to fully appear.[4]

Section 2 explains the theory of Benford's Law, derives the distribution, and gives an introduction to the literature. Obviously Benford's Law is an "oddball" statistical regularity, which seems unknown to most economists. However, as argued by Varian (1972), it points to a problem if it fails. It may indicate that something "fishy" is going on.

Section 3 explains the calculations from the gdp-matrix to the first digit distribution and shows how the resulting distributions look. It is also analyzed how close the distribution is to Benford's distribution. It fits amazingly well, but not perfectly well. The calculations are also made for different periods and for different country groups. The paper looks at all data, and at two quartile distributions: Over time in section 4 and over gdp-levels in section 5.

2. The data were collected by the late Angus Maddison for the OECD Millennium Publication (see references). Maddison updated the data till February 2010 (as used at present), shortly before he passed away.
3. The published data has no decimals, but the downloadable worksheets have plenty. When I refer to # digits, it is the digits before the decimal point, i.e., to the form of the published data.
4. Imagine, e.g., that we consider a sub-sample of gdp observations that all are in the range from $1000 to 1800. Then the fist digit would be 1 for all observations, and no Benford distribution would appear. Thus, it is a condition for the Benford distribution to appear that the range of the data is large enough.

In addition to Benford's Law, we have reason to consider the folk theorem in empirical economics: *If you test any theory in economics with enough data, you will reject it.* Benford's Law is not an economic theory – it is rather a numerical regularity, but nevertheless the results turn out to be that the law seems to apply rather well, but it is rejected by a formal test using all the data.

2. Why should Benford's Law apply?

Benford's Law is one of these obvious relations that you are unlikely to see for yourself – and the first time it is presented it looks like a conjuring trick. The whole point is obvious once you look at a logarithmic axis: The distance between 1 and 2 is much larger than the distance between 2 and 3 etc., so in a log-linear world more numbers start with 1 than with 2!

It is easy to calculate the frequencies of Table 1 by noting that log(1) = 0.000 and log(2) = 0.3010, so the distance between the two is 0.3010, the distance between log(2) and log(3) is 0.1761, etc. And, lo and behold, the sum of all distances between log(1) and log(10) is 1, as follows from the fact that log(10) = 1. Hence by multiplication by 100 we get the frequencies in % of Benford's distribution from Table 1.

Benford (1938) describes the complicated path that led him to his law. He believed that he was dealing with "anomalous numbers". Also, he did find some rather puzzling data sets that obeyed the law. Varian (1972) proposes to use the law to reveal data manipulation, and Nigrini (1996 and 1999) and Nye and Mould (2007) study cases where the data do obey the law and others where the law fails under suspicious circumstances.

The newest paper in the field is Michalski and Stoltz (2010) dealing with balance of payment data. For most countries these data obey Benford's Law. However, countries with fixed exchange rates under pressure publish data that does not follow the law at all. Michalski and Stoltz interpreted this as a clear statistical proof that countries under pressure manipulate their data. Not all countries do it, but enough to generate highly suspicious data.

The data in the Maddison set should be log-linear, and they are put together by many agencies and individual researchers. Thus, theoretically Benford's Law should fit.

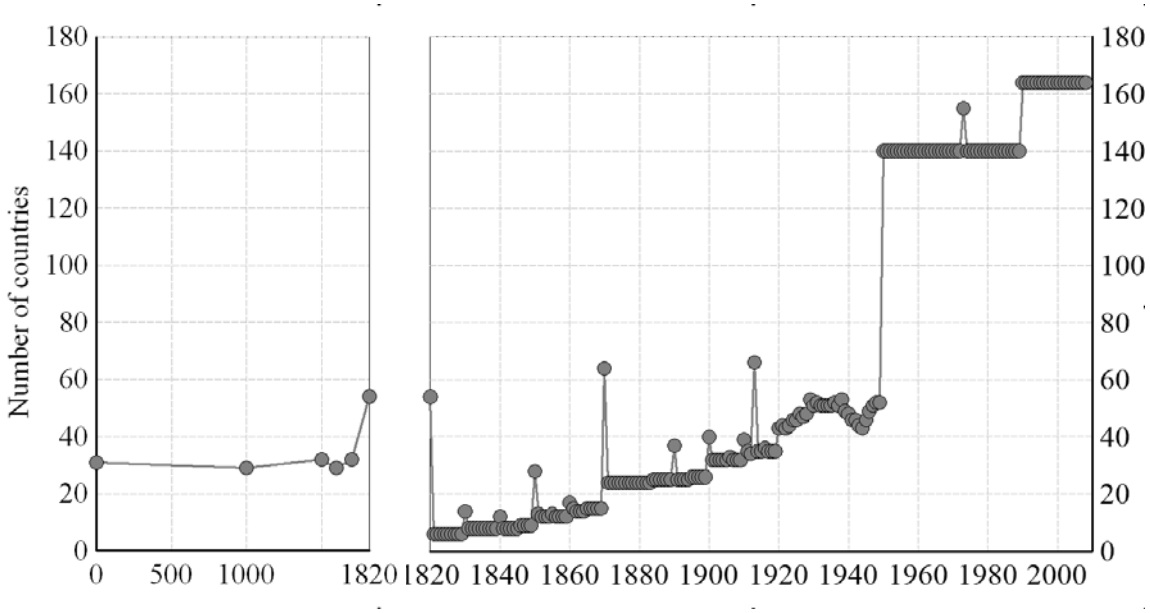
3. Looking at all 12,411 gdp data in the Maddison set

Section 3.1 shows the coverage of the data. Section 3.2 shows how the calculations are made for all observations and discusses the truncation problem. Section 3.3 gives the formal tests of the distributions.

3.1 The coverage of the data

As of now the Maddison gdp data consists of 12,411 observations. They cover – as much as possible – the period from year 1 to 2008 for 164 (present day) countries. Figure 1 shows the number of countries covered. The first 1800 years had only a few observations, but gradually more countries are included.

Figure 1. The number of observations in the Maddison gdp data set



Note: The left-hand part of the figure ends in 1820, where the right-hand part starts. The dots are for the years with data and show the number of countries covered. Observations are available for 2008 – 1820 + 6 = 194 years, and 164 countries; but of the full 194 *x* 164 matrix, only 12,411/31,816 = 39% of the cells are filled.

3.2 The calculations

The calculations are done in a standard worksheet and use the histogram function:

The minimum gdp is 207 in the Maddison data set, and the maximum is 42,196; hence they extend over three log-decades. Eight are from the log-decade of the hundreds: 200, 300, 400 500, 600, 700, 800 and 900; nine are from the log-decade of the thousands 1000, 2000,

3000, 4000, 5000, 6000, 7000, 8000 and 9000; and four are from the log-decade of the ten thousands 10000, 20000, 30000, 40000. When the bin limits in the histogram are set at the said numbers, it generates the first-digit-frequencies listed in Table 2.[5] The last column to the right is the sums of the frequencies starting with each digit. They are the frequencies that should follow Benford's Law.

Table 2. The basic count: Number of observations in three log decades

First digit	All observations between 207 and 42,196			
	Hundreds	Thousands	Ten thousands	All
1		2988	945	3933
2	15	1758	271	2044
3	66	1205	37	1308
4	268	781	1	1050
5	449	531		980
6	586	384		970
7	476	300		776
8	450	240		690
9	482	178		660
Sum	2792	8365	1254	12411

Note: The full panel has 194·164 = 31,816 cells, of which 12,411 are filled.

The same procedure can be used for any subset of the data. It is easy to check that all data are counted once and only once as the sum of all parts of any division has to tally with the values given in Table 2. As mentioned there are no observations below 200 and none above 50,000, and the number of observations become very thin when they approach the two ends of the interval. So measured in log-decades, the range of the data seems barely enough. Figure 2 shows the observations from Table 2, and includes the distribution of all observations expected from Benford's Law.

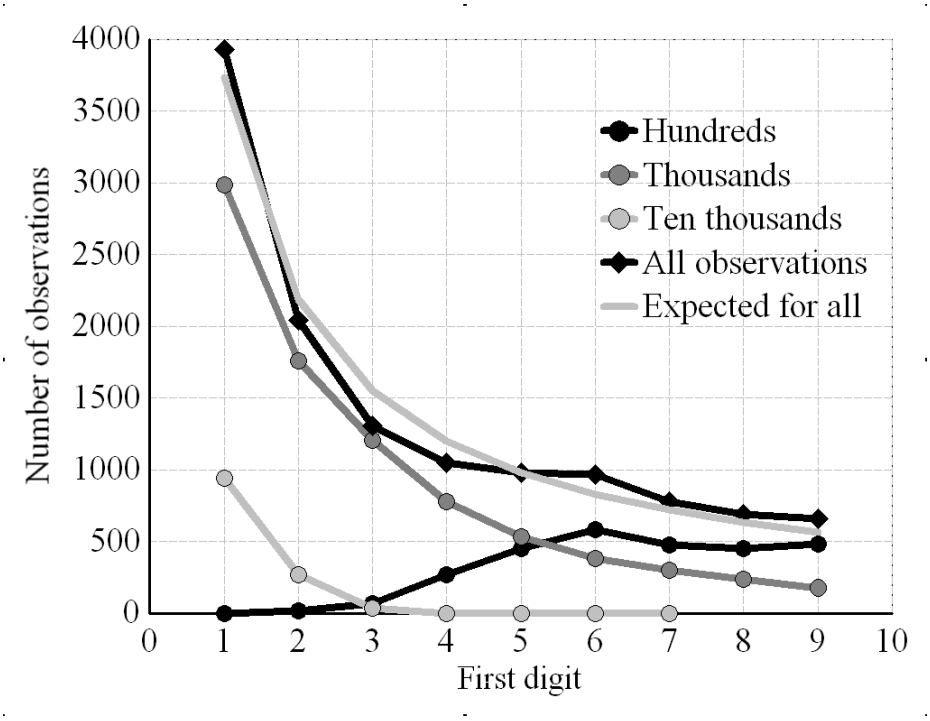
From just looking at the curves, it is obvious that the log-decades of the hundreds and the ten thousands do not follow the Benford distribution – they are too close to the start and the end of the data. The frequencies for the log-decade of the hundreds only look Benford-like after the fifth digit. And the frequencies for the log-decade of the ten-thousands quickly fall to zero. Thus the distribution is likely to suffer from some range problems.

However, both the curve for all observations and the one for the log-decades of the thousands look much as expected. The impression one gets from the graph is that the

5. To guard against rounding errors, all bins were set 0.0001 lower.

frequencies for the thousands fall too fast as they get close to the end. Also, the curve for all has an extra hump at 6, when the curve for the hundreds kicks in.

Figure 2. The frequencies from Table 3 graphed,
with the Benford distribution of 12,411 observations included



2.3 The formal tests: Nothing fits perfectly

The formal tests of the two most Benford-looking frequencies from Table 2 above are given in Table 3. They confirm the observations from Figure 2. For all observations there are too few observations for digits 2 and 3 and too many for digit 6. For the log-decade of the thousands, the distribution is too steep.

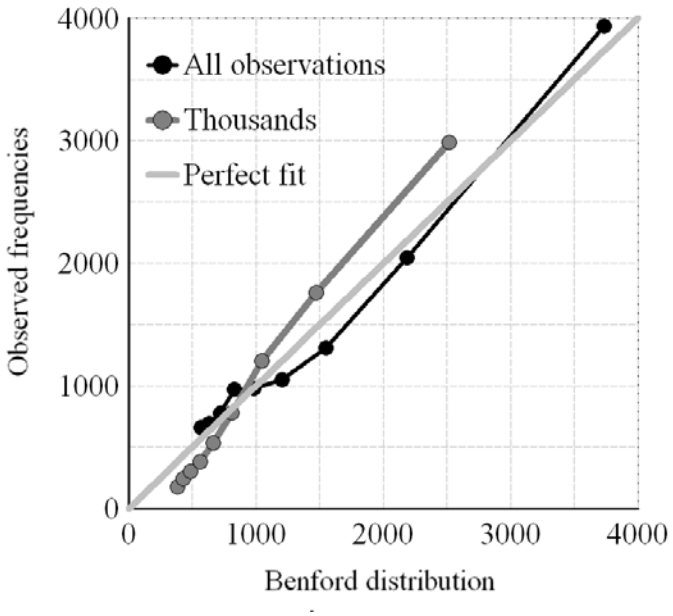
However, as is shown on Figure 3 the observed and the expected distributions do look alike. It is not likely that any other known distribution will have a better fit, and we have a very simple theory to explain why it should fit. Also, if we move 300 of the 12,411 observations in the right way, we can get the χ^2-test within the 5% limit of significance, and if we move about 100 more, we can reach a χ^2-test at the 50% level.

Table 3. Two formal tests

	All observations				The observations of the ten thousands			
Digit	Observed	Expected	Diff.	χ^2-test	Observed	Expected	Diff.	χ^2-test
1	3,933	3,736.1	196.9	**10.4**	2988	2518.1	469.9	**87.7**
2	2,044	2,185.5	-141.5	**9.2**	1758	1473.0	285.0	**55.1**
3	1,308	1,550.6	-242.6	**38.0**	1205	1045.1	159.9	**24.5**
4	1,050	1,202.8	-152.8	**19.4**	781	810.7	-29.7	1.1
5	980	982.7	-2.7	0.0	531	662.4	-131.4	**26.0**
6	970	830.9	139.1	**23.3**	384	560.0	-176.0	**55.3**
7	776	719.7	56.3	4.4	300	485.1	-185.1	**70.6**
8	690	634.9	55.1	4.8	240	427.9	-187.9	**82.5**
9	660	567.9	92.1	**14.9**	178	382.8	-204.8	**109.5**
Sum	12,411	12,411	0	**124.3**	8365	8365	0	**512.4**

Note: All χ^2-test contributions of 8 or higher are bolded.

Figure 3. A scatter of observed over expected frequencies



Thus we can explain the deviations from Benford's distribution by the range problems, and it is arguable that the fit is as good as it can be under the circumstances.

4. Two divisions into quartiles: By time and by income

In the two division experiments, the 12,411 observations are divided in quartiles. Each quartile has to be as close to 12,411/4 = 3,103.75 observations as possible. In section 4.1 the quartiles are made over time, so that each year is kept together. In section 4.2 the quartiles are made by income, so that each country is kept together.

4.1 Division by time

It is possible that the problems are concentrated in one end of the data, and on the face of it one may suspect that the old data are the most problematic. Table 4 shows the division made and the test results for each quartile.

Table 4. Four quartiles over time

	1st quartile			2nd quartile			3rd quartile			4th quartile		
Years	From 1 to 1937			From 1938 to 1967			From 1968 to 1989			From 1990 to 2008		
Digit	N	Diff.	χ^2-test	N	Diff.	χ^2-test	N	Diff	χ^2-test	N	Diff.	χ^2-test
1	1100	166.5	**29.7**	889	-43.9	2.1	1018	86.3	**8.0**	926	-12.0	0.2
2	599	52.9	5.1	500	-46.7	3.8	378	-167.0	**51.2**	567	18.3	0.6
3	370	-17.4	0.8	337	-52.2	6.5	283	-103.7	**27.8**	318	-71.3	**13.1**
4	290	-10.5	0.4	262	-38.3	4.9	263	-36.9	4.5	235	-67.0	**14.9**
5	192	-53.5	**11.7**	254	8.6	0.3	266	20.9	1.8	268	21.3	1.8
6	162	-45.6	**10.0**	283	74.5	**27.5**	278	70.8	**24.2**	247	38.4	7.1
7	130	-49.8	**13.8**	228	50.3	**13.0**	194	14.5	1.2	224	43.3	**10.4**
8	121	-37.6	**8.9**	161	1.5	0.0	215	56.7	**20.3**	193	33.6	7.1
9	137	-4.9	0.2	185	47.2	**13.2**	200	58.4	**24.1**	138	-4.6	0.1
Sum	3101	-0	80.6	3099	0	71.3	3095	0	163.0	3116	0	55.2

Note: The table is calculated as Table 3, but the column for the expected distribution is omitted.

Table 5. Kendall's τ between the nine χ^2-contributions from Tables 3 and 4

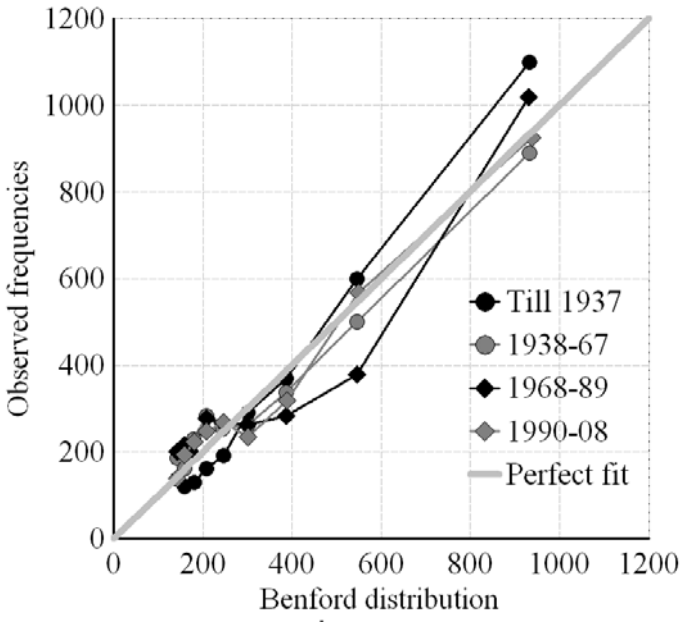
	1st quartile	2nd quartile	3rd quartile	4th quartile	All	Trend [b]
1st quartile	1	-0.22	-0.33	-0.17	-0.33	-0.22
2nd quartile	-0.22	1	0.11	-0.06	0.44	0.22
3rd quartile	-0.33	0.11	1	-0.17	0.44	-0.22
4th quartile	-0.17	-0.06	-0.17	1	0.06	0.06
Average [a]	-0.24	-0.06	-0.13	-0.13	0.15	-0.04

Note: a) Average of non diagonal elements. b) Correlation to the nine digits. None of the τ's calculated are significant.

Three quartiles have lower χ^2-tests than the aggregate one, and all reject that the data for the quartile follow Benford's Law. The quartile that deviates most from the Benford distribution is the 3rd one that goes from 1968 to 1989.

Once again, the frequencies for each quarter do not look very different from the expected distribution as seen on Figure 4. Also the largest χ^2-contributions are for different digits. Table 5 shows the correlations between the nine χ^2-contributions for the four quartiles. The correlations are calculated by Kendall τ, chosen as the χ^2-contributions are squared. The correlations in Table 5 are not very high, and the impression that the largest χ^2-contributions are fairly randomly distributed is confirmed by Figure 4.

Figure 4. A scatter of observed over expected frequencies for the time-quartiles



We are thus in the same situation as in section 2. The four quartiles look much as they would if they followed Benford's Law, but the formal test rejects that the deviations are random.

4.2 Divided by income

The second experiment divides the countries by income level. Everything except the division is done precisely as in section 4. The division follows the present gdp level. Thus the gdp-matrix is first sorted by the average income for the last decade, which is for the period from 1999 to 2008, and then the countries are divided in 4 groups as explained.

Table 6. Four quartiles over countries by income

Income	1st quartile Lowest			2nd quartile Lower middle			3rd quartile Higher middle			4th quartile Highest		
Digit	N	Diff.	χ^2-test	N	Diff.	χ^2-test	N	Diff	χ^2-test	N	Diff.	χ^2-test
1	1149	209.5	**46.7**	812	-124.5	**16.6**	888	-47.3	2.4	1084	159.2	**27.4**
2	140	-409.6	**305.2**	752	204.2	**76.1**	457	-90.1	**14.8**	695	154.0	**43.9**
3	80	-309.9	**246.3**	543	154.3	**61.3**	320	-68.2	**12.0**	365	-18.8	0.9
4	169	-133.5	**58.9**	301	-0.5	0.0	294	-7.1	0.2	286	-11.7	0.5
5	337	89.9	**32.7**	189	-57.3	**13.3**	277	31.0	3.9	177	-66.2	**18.0**
6	400	191.1	**174.7**	190	-18.3	1.6	254	46.0	**10.2**	126	-79.7	**30.9**
7	299	118.0	**76.9**	131	-49.4	**13.5**	231	50.8	**14.3**	115	-63.2	**22.4**
8	275	115.4	**83.3**	81	-78.1	**38.4**	216	57.1	**20.5**	118	-39.1	**9.7**
9	272	129.2	**116.9**	112	-30.4	6.5	170	27.8	5.4	106	-34.6	8.5
Sum	3121	0	**1141.7**	3111	0	**227.2**	3107	0	**83.7**	3072	0	**162.2**

Note: See Table 4.

Table 7. Kendall's τ between the nine χ^2-contributions from Tables 3 and 6

	1st quartile	2nd quartile	3rd quartile	4th quartile	All	Trend [b)]
1st quartile	1	0.33	0.44	0.11	0.44	0.00
2nd quartile	0.33	1	0.44	0.22	-0.11	-0.22
3rd quartile	0.44	0.44	1	0.22	-0.11	0.22
4th quartile	0.11	0.22	0.22	1	-0.22	-0.22
Average [a)]	0.30	0.33	0.37	0.19	0.00	-0.06

Here the deviations from Benford's distribution are larger, and the largest deviation occurs for the poorest quarter of the countries, where the $\chi^2(8) = 1141.7$. This is a dramatic rejection, and on Figure 4 it is certainly clear that the curve for the first quartile deviates a lot from the expected one. However, when we look at all four curves the deviations are very different, as they have to be when the average curve looks as Figure 2.

The countries in the highest group were once poor, but the countries in the lowest group were never rich. Thus the range problems are likely to be concentrated in the lowest group. This is precisely what is shown by Figure 6, which is made as Figure 2. The curves are a bit difficult to read as they overlap. The frequencies are from two incomplete log-decades. In the 100s the Benford distribution first starts to appear from digit 6 and up, then the frequencies jump to the 1000s, but here the frequencies very quickly taper off. This explains why the aggregate distribution misses the observations from 2 to 5. In this case the range problem is so large that it dominates the data.

Figure 5. A scatter of observed over expected frequencies for the income-quartiles

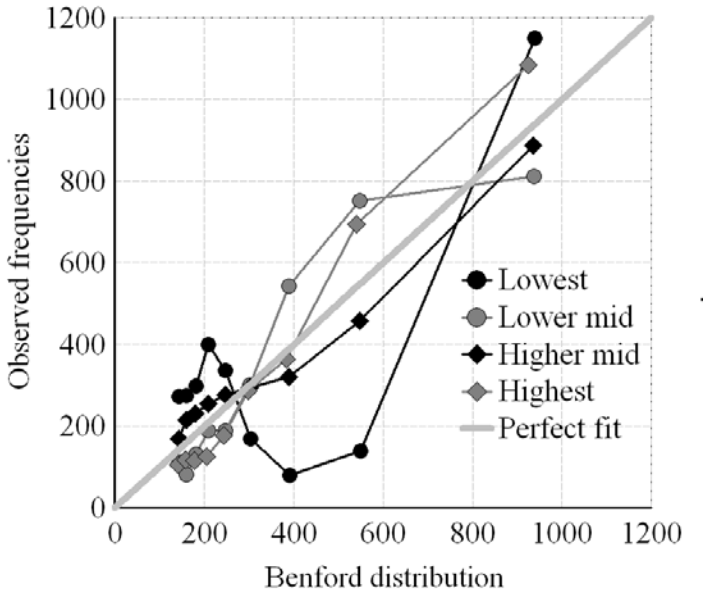
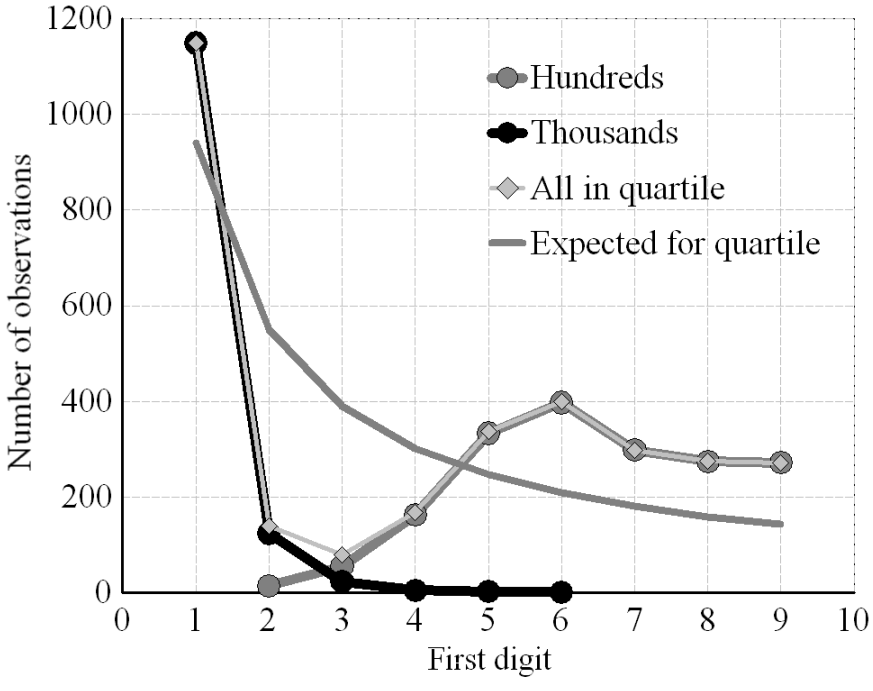


Figure 6. The frequencies from the lowest quartile in Table 6 graphed.

The Benford distribution of 3,121 observations included



6. How worried should we be?

The author of this paper is a frequent user of the Maddison gdp-data both in his research and his teaching. Many other economists use these data, and hence it appears a worthwhile project to submit them to a test. Consequently we study if they follow Benford's Law. The only economics it demands is that it assumes that data for gdp (GDP per capita) have a basic log-linear structure.

Our analysis has showed three points: (1) The first-digit data has a frequency structure that looks rather close to Benford's distribution. (2) A formal test shows that the deviations from the distribution are too large to be random. (3) When the data are divided in groups it improves the fit when the division is over time, but when the division is over income the fit gets worse especially in the quartile for the lowest income.

As regards (2) the paper contains 10 formal tests to determine if the Maddison gdp-data follow Benford's Law, either for the whole set or for various subdivisions. In all ten cases, the tests reject that the distribution is an exact description of the data. If we follow Varian (1972), we should conclude that the data are manipulated, but it is not easy to point to a manipulator with some purpose in mind. These data are, for all I know, a careful compilation of almost everything we know about the long-run trends in economic development.

Hence, I propose that the deviations from the law are due to range problems. The data cover only three log-decades, the hundreds, the thousands and the ten-thousands, and two of these log-decades are not fully covered. This gives some minor kinks in the frequency-curve which causes the test to reject that the law is perfectly fulfilled. When the countries are divided by income, the range problems are concentrated in the low income group, where the law is dramatically rejected.

References:

Benford, F. 1938. The law of anomalous numbers. *Proceedings of the American Philosophy Society* 78, 551-72

Maddison homepage: http://www.ggdc.net/maddison/

Maddison, A., 2001. *The world economy: A millennial Perspective.* OECD, Paris

Maddison, A., 2003. *The world economy: Historical statistics.* OECD, Paris

Michalski, T., Stoltz, G., 2010. Do countries falsify economic data strategically? Some evidence that they do. Paper presented at Degit 2010

Nigrini, M., 1996. Taxpayer compliance application of Benford's law. *Journal of the American Taxation Association* 18, 72-92

Nigrini, M., 1999. I've got your number: How a mathematical phenomenon can help CPAs uncover fraud and other irregularities. *Journal of Accountancy*, May

Nye, J., Moul, C., 2007. The political economy of numbers: on the application of Benford's law to international macroeconomic statistics. *The B.E. Journal of Macroeconomics* 7 (1), article 17

Varian, H.R., 1972. Benford's Law. *The American Statistician* 26(3), 65-6