OCR
RECOGNISING ACHIEVEMENT

ADVANCED GCE UNIT

4754(B)/01

MATHEMATICS (MEI)

Applications of Advanced Mathematics (C4)

Paper B: Comprehension

INSERT

TUESDAY 23 JANUARY 2007

Afternoon
Time: Up to 1 hour

INSTRUCTIONS TO CANDIDATES

- This insert contains the text for use with the questions.

This document consists of **10** printed pages and **2** blank pages.

Benford's Law

Leading digits

This article is concerned with a surprising property of the leading digits of numbers in various sets. The leading digit of a number is the first digit you read. In the number 193 000 the leading digit is 1. When a number is written in standard form, such as 1.93×10^5 or 2.78×10^{-7}, the leading digit is the digit before the decimal point, in these examples 1 and 2 respectively.

Mathematical sequences

Table 1 shows the integer powers of 2, from 2^1 to 2^{50}. In this table, which digits occur more frequently as the leading digit? You might expect approximately one ninth of the numbers to have a leading digit of 1, one ninth of the numbers to have a leading digit of 2, and so on. In fact, this is far from the truth.

2	2 048	2 097 152	2 147 483 648	2 199 023 255 552
4	4 096	4 194 304	4 294 967 296	4 398 046 511 104
8	8 192	8 388 608	8 589 934 592	8 796 093 022 208
16	16 384	16 777 216	17 179 869 184	17 592 186 044 416
32	32 768	33 554 432	34 359 738 368	35 184 372 088 832
64	65 536	67 108 864	68 719 476 736	70 368 744 177 664
128	131 072	134 217 728	137 438 953 472	140 737 488 355 328
256	262 144	268 435 456	274 877 906 944	281 474 976 710 656
512	524 288	536 870 912	549 755 813 888	562 949 953 421 312
1 024	1 048 576	1 073 741 824	1 099 511 627 776	1 125 899 906 842 624

Table 1

The frequencies of the different leading digits in these powers of 2 are shown in Table 2.

Leading digit	1	2	3	4	5	6	7	8	9
Frequency	15	10	5	5	5	4	1	5	0

Table 2

You can see that, for these data, 1 and 2 appear more frequently than any other numbers as the leading digit. Is this just a peculiarity of the first fifty powers of 2, or is a general pattern emerging?

Here is another example. Imagine you invest £100 in an account that pays compound interest at a rate of 20% per year. Table 3 shows the total amount (in £), after interest is added, at the end of each of the following 50 years.

120.00	743.01	4 600.51	28 485.16	176 372.59
144.00	891.61	5 520.61	34 182.19	211 647.11
172.80	1 069.93	6 624.74	41 018.63	253 976.53
207.36	1 283.92	7 949.68	49 222.35	304 771.83
248.83	1 540.70	9 539.62	59 066.82	365 726.20
298.60	1 848.84	11 447.55	70 880.19	438 871.44
358.32	2 218.61	13 737.06	85 056.22	526 645.73
429.98	2 662.33	16 484.47	102 067.47	631 974.87
515.98	3 194.80	19 781.36	122 480.96	758 369.85
619.17	3 833.76	23 737.63	146 977.16	910 043.82

Table 3

The frequencies of leading digits for these data are shown in Table 4.

Leading digit	1	2	3	4	5	6	7	8	9
Frequency	15	9	6	5	4	3	4	2	2

Table 4

The pattern that emerges is much the same as that in Table 2, with 1 appearing as the leading digit in 30% of cases and 2 appearing in 18% of cases.

Now imagine that a second person invests £200 rather than £100. Each amount in this person's table (Table 5) is double the corresponding amount in Table 3.

240.00	1 486.02	9 201.02	56 970.32	352 745.18
288.00	1 783.22	11 041.23	68 364.38	423 294.21
345.60	2 139.86	13 249.47	82 037.25	507 953.05
414.72	2 567.84	15 899.37	98 444.70	609 543.66
497.66	3 081.40	19 079.24	118 133.65	731 452.40
597.20	3 697.69	22 895.09	141 760.37	877 742.88
716.64	4 437.22	27 474.11	170 112.45	1 053 291.45
859.96	5 324.67	32 968.93	204 134.94	1 263 949.74
1 031.96	6 389.60	39 562.72	244 961.93	1 516 739.69
1 238.35	7 667.52	47 475.26	293 954.31	1 820 087.63

Table 5

The frequencies of leading digits for these data are shown in Table 6.

Leading digit	1	2	3	4	5	6	7	8	9
Frequency	15	9	6	5	4	3	3	3	2

Table 6

A remarkable result now emerges. The frequencies in Table 6 are almost the same as those in Table 4.

In Table 3 there are 15 numbers with a leading digit of 1. Each of these numbers, when doubled, has a leading digit of either 2 or 3, as can be seen in Table 5. Similarly, the numbers in Table 3 with leading digit 5, 6, 7, 8 or 9 give numbers in Table 5 with leading digit 1. These outcomes are reflected in the frequencies in Table 4 and Table 6.

Physical phenomena

The numbers in Tables 1, 3 and 5 were all generated mathematically. Now look at something less mathematical in origin.

The frequencies of the leading digits of the areas of the world's 100 largest countries, measured in square kilometres, are given in Table 7. (For interest the data are given in Appendix A.)

Leading digit	1	2	3	4	5	6	7	8	9
Frequency	33	20	13	8	6	6	4	4	6

Table 7

You will notice that these data, even though they have a non-mathematical origin, show essentially the same pattern of frequencies. The populations of cities and countries also show this general pattern. As further examples, if you take the heights of the world's tallest mountains, the lengths of Europe's longest rivers, the numbers of votes cast for each political party in every constituency in a general election or the values of a wide range of scientific constants, you will find a similar pattern in many cases. The remainder of this article looks at such physical data, rather than mathematically generated data, and answers the following question.

Why does this pattern in leading digits occur, and how can it be modelled mathematically?

Benford's Law

This phenomenon was noted in 1881 by Simon Newcombe, an American mathematician and astronomer, and then rediscovered by the physicist Frank Benford in 1938. Benford analysed 20 229 sets of data, including information about rivers, baseball statistics and all the numbers in an issue of *Reader's Digest*. He was rewarded for his efforts by having the law named after him.

Benford's Law gives a formula for the proportions of leading digits in data sets like these. This formula will be derived over the next few pages.

The proportions given by Benford's Law are illustrated in Fig. 8.

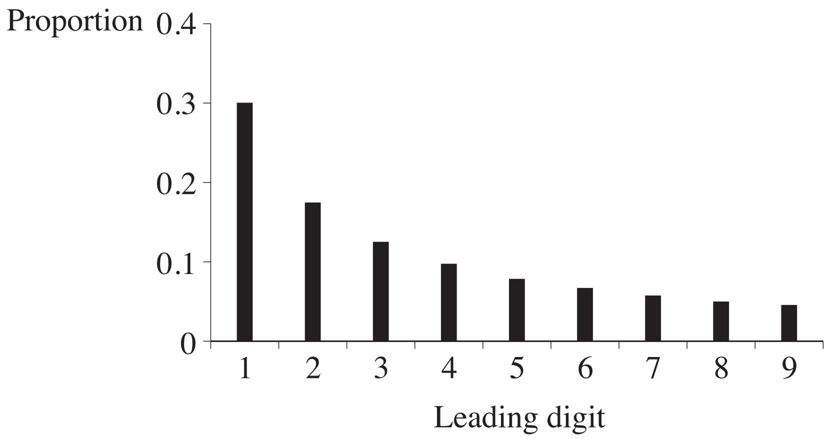

Fig. 8

This shows that, in a typical large data set, approximately 30% of the data values have leading digit 1 but fewer than 5% have leading digit 9. For small data sets, you cannot expect the leading digits to follow Benford's Law closely; the larger the data set, the better the fit is likely to be.

Fig. 9 shows the proportions for the leading digits of the areas of the world's largest countries, derived from Table 7, together with the proportions given by Benford's Law. You will see that there is a very good match.

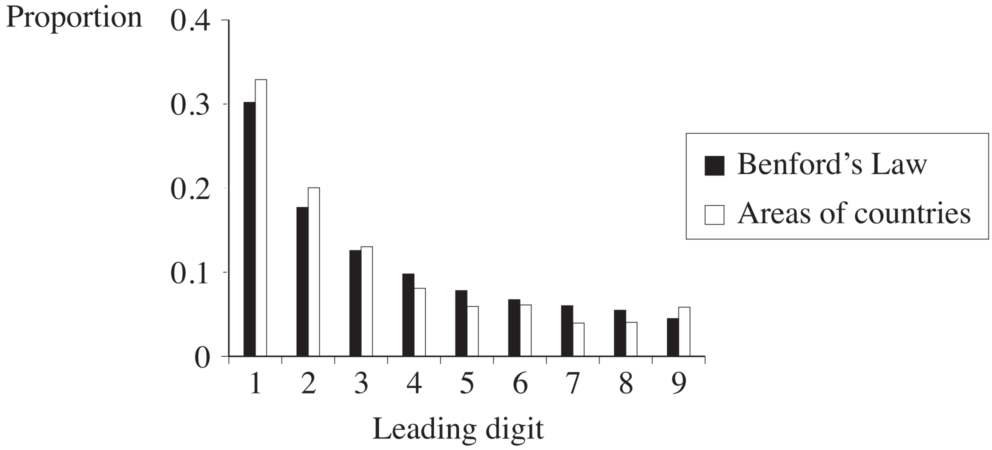

Fig. 9

Benford's Law does not apply to all situations, even when there is a large data set. There is still debate about the conditions under which it applies. The rest of this article relates to situations in which it does apply.

Scale invariance

If the areas of the countries in Appendix A are measured in square miles, rather than square kilometres, it turns out that the leading digits still follow the same pattern. This is a feature of all data sets which follow Benford's Law; it does not matter what units are used when measuring the data. The next example illustrates this.

Table 10 shows the frequencies of the leading digits of the share prices of the 100 largest UK companies on 28 February 2006, in pounds sterling, US dollars and euros.

Leading digit	1	2	3	4	5	6	7	8	9
Frequency (£)	33	14	7	4	14	11	3	8	6
Frequency ($)	41	18	13	6	6	2	3	7	4
Frequency (€)	36	18	10	8	2	4	8	8	6

Table 10

These results are illustrated in Fig. 11 along with the proportions given by Benford's Law. Despite there being only 100 items of data, two features are evident.

- There is a reasonable agreement between the proportions for the three currencies.
- Benford's Law gives a reasonable approximation in each case.

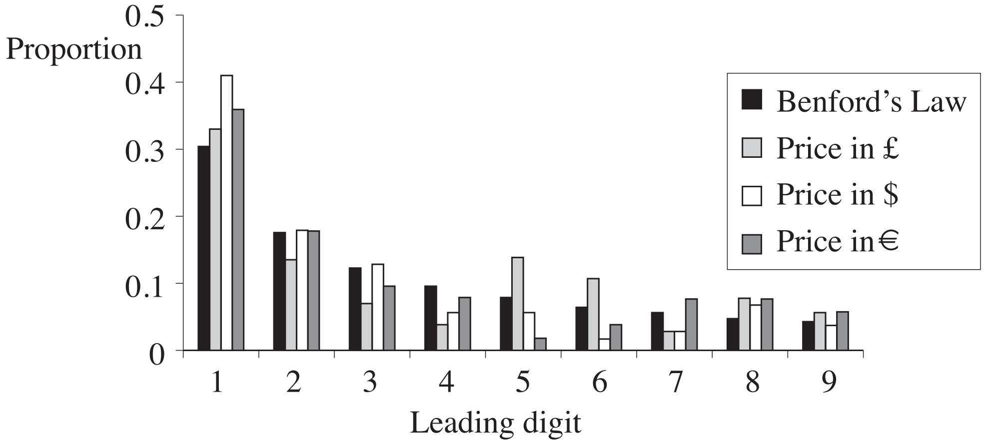

Fig. 11

This property, that it does not matter what units are used when measuring the data, is called *scale invariance*.

The frequencies of leading digits

The idea of scale invariance is important. If scale invariance applies, what does this tell us about the frequencies of leading digits?

In order to answer this question it is helpful to use the following notation.

- p_n represents the proportion of data values with leading digit n.

Thus p_1 represents the proportion of data values with leading digit 1, p_2 represents the proportion of data values with leading digit 2, and so on. Clearly $\sum_{n=1}^{9} p_n = 1$.

The proportions $p_1, p_2, \ldots, p_9$ can be represented by the areas of the rectangles on a diagram such as Fig. 12. The total area is 1.

p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8	p_9

Fig. 12

Many things can be deduced about the values of $p_1, p_2, \ldots, p_9$ by thinking about a large data set in which scale invariance holds exactly. Here are some of them.

- If every number in the data set is multiplied by 2, then all the numbers with leading digit 1, and no others, are mapped to numbers with leading digit either 2 or 3. Since this does not change the distribution of leading digits, it follows that

$$p_1 = p_2 + p_3.$$

- Similarly, when multiplying by 2, all numbers with leading digit 5, 6, 7, 8 or 9, and no others, are mapped to numbers with leading digit 1. Therefore

$$p_5 + p_6 + p_7 + p_8 + p_9 = p_1.$$

As a consequence of these two results,

$$p_1 + (p_2 + p_3) + p_4 + (p_5 + p_6 + p_7 + p_8 + p_9) = 3p_1 + p_4,$$

from which it follows that

$$3p_1 + p_4 = 1$$

and so $p_1 < \frac{1}{3}$.

- By using a multiplier of 4, instead of 2, it follows that

$$p_1 = p_4 + p_5 + p_6 + p_7.$$

This shows that $p_1 > p_4$. Using the fact that $3p_1 + p_4 = 1$, it follows that $p_1 > \frac{1}{4}$. Therefore $\frac{1}{4} < p_1 < \frac{1}{3}$. This is consistent with the value of about 0.3 observed in several of the data sets considered earlier in the article.

In a similar way, other relationships connecting values of p_n, such as $p_1 = p_3 + p_4 + p_5$, $p_6 + p_7 = p_3$ and $p_2 = p_6 + p_7 + p_8$, can be derived.

Deriving Benford's Law

It is helpful now to introduce the quantities L(1), L(2), …, L(10), defined as follows.

- $\mathrm{L}(1) = 0$
- $\mathrm{L}(2) = p_1$
- $\mathrm{L}(3) = p_1 + p_2$
- $\mathrm{L}(4) = p_1 + p_2 + p_3$

  …
- $\mathrm{L}(10) = p_1 + p_2 + \ldots + p_9 = 1$

The quantities L(1), L(2), …, L(10) are the cumulative proportions. They are illustrated in Fig. 13.

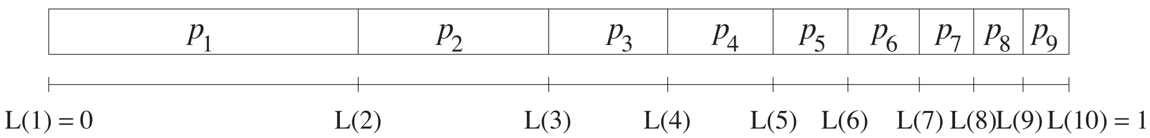

Fig. 13

What can you say about the quantities L(1), L(2), …, L(10)?

- You know $p_1 = p_2 + p_3$.

  This corresponds to $\mathrm{L}(2) - \mathrm{L}(1) = \mathrm{L}(4) - \mathrm{L}(2)$ which simplifies to $\mathrm{L}(4) = 2 \times \mathrm{L}(2)$.

- Similarly $p_1 = p_3 + p_4 + p_5$.

  This corresponds to $\mathrm{L}(2) - \mathrm{L}(1) = \mathrm{L}(6) - \mathrm{L}(3)$ which simplifies to $\mathrm{L}(6) = \mathrm{L}(3) + \mathrm{L}(2)$.

- Also $p_6 + p_7 = p_3$.

  This corresponds to $\mathrm{L}(8) - \mathrm{L}(6) = \mathrm{L}(4) - \mathrm{L}(3)$. Combining this with the last two results gives $\mathrm{L}(8) = 3 \times \mathrm{L}(2)$.

These results, and others like them, suggest that $\mathrm{L}(n)$ is a logarithmic function. The fact that $\mathrm{L}(10) = 1$ shows that the base of the logarithms is 10, and so $\mathrm{L}(n) = \log_{10} n$.

It follows that $p_n = \mathrm{L}(n+1) - \mathrm{L}(n) = \log_{10}(n+1) - \log_{10} n$. That is, the proportion of data values with leading digit n (where $1 \leqslant n \leqslant 9$) is $\log_{10}(n+1) - \log_{10} n$. This is Benford's Law.

Uses of Benford's Law

Since the 1980s Benford's Law has, on several occasions, been used successfully to convict people accused of fraud. When concocting figures to include in fictitious company accounts, it is natural to try to make the amounts look 'random' or 'average'. This might, for example, be done by including a high proportion of 'average' amounts beginning with 'middle digits' such as 4, 5 or 6, or including amounts just under £1000, £10 000 and £100 000, in an attempt to avoid closer analysis. In this way fraudsters are generating data which do not follow Benford's Law and thereby attracting the kind of scrutiny they were trying to avoid.

Appendix A Areas of countries (thousands of km^2)

Russia	17 075	Pakistan	804	Poland	313
Canada	9976	Mozambique	802	Italy	301
USA	9629	Turkey	781	Philippines	300
China	9597	Chile	757	Ecuador	284
Brazil	8512	Zambia	753	Burkina Faso	274
Australia	7687	Myanmar	679	New Zealand	269
India	3288	Afghanistan	648	Gabon	268
Argentina	2767	Somalia	638	Western Sahara	266
Kazakhstan	2717	C. African Republic	623	Guinea	246
Sudan	2506	Ukraine	604	Great Britain (and N Ireland)	245
Algeria	2382	Botswana	600	Ghana	239
Congo (Dem. Rep.)	2345	Madagascar	587	Romania	238
Greenland	2166	Kenya	583	Laos	237
Mexico	1973	France	547	Uganda	236
Saudi Arabia	1961	Yemen	528	Guyana	215
Indonesia	1919	Thailand	514	Oman	212
Libya	1760	Spain	505	Belarus	208
Iran	1648	Turkmenistan	488	Kyrgyzstan	199
Mongolia	1565	Cameroon	475	Senegal	196
Peru	1285	Papua New Guinea	463	Syria	185
Chad	1284	Sweden	450	Cambodia	181
Niger	1267	Uzbekistan	447	Uruguay	176
Angola	1247	Morocco	447	Tunisia	164
Mali	1240	Iraq	437	Suriname	163
South Africa	1220	Paraguay	407	Bangladesh	144
Colombia	1139	Zimbabwe	391	Tajikistan	143
Ethiopia	1127	Japan	378	Nepal	141
Bolivia	1099	Germany	357	Greece	132
Mauritania	1031	Congo (Rep.)	342	Nicaragua	129
Egypt	1001	Finland	337	Eritrea	121
Tanzania	945	Malaysia	330	Korea (North)	121
Nigeria	924	Vietnam	330	Malawi	118
Venezuela	912	Norway	324		
Namibia	825	Cote d'Ivoire	322		

BLANK PAGE

BLANK PAGE