Frequency of letters in the English langauge

Dr Richard Kenderdine

It is well-known that significant differences occur in the frequency of the letters in the English langauge. Table 1 shows the relative frequencies of the letters contained in a number of magazines and novels, originally published in *Cipher Systems:The Protection of Communication.* The sample size was 100 362. This distribution is used for comparison and is referred to as the standard distribution.

Letter	a	b	c	d	e	f	g	h	i	j	k	l	m
Percent	8.2	1.5	2.8	4.3	12.7	2.2	2	6.1	7	0.2	0.8	4.0	2.4
Letter	n	o	p	q	r	s	t	u	v	w	x	y	z
Percent	6.7	7.5	1.9	0.1	6	6.3	9.1	2.8	0.1	2.4	0.2	2.0	0.1

Table 1: Standard percentage distribution of letters in the English language

This note has two purposes:

(1) To compare the relative frequencies shown in Table 1 with those obtained from four different texts in diverse fields of writing: novel, religious, political and scientific.

(2) To contrast the result from one text with the outcome after applying a cipher.

Four texts from different fields of writing

The initial step before analysing the texts is to convert all letters to upper case. The clearest way to show the comparison is to use a side-by-side bar chart.

The first text is the novel ‘Pride and Prejudice’ with 536 378 letters. The bar chart in Figure 1 indicates that there is not much difference in the distribution of letters comparing ‘Pride and Prejudice’ with the standard, except for the greater use of the letter ‘*v*’ in the former.

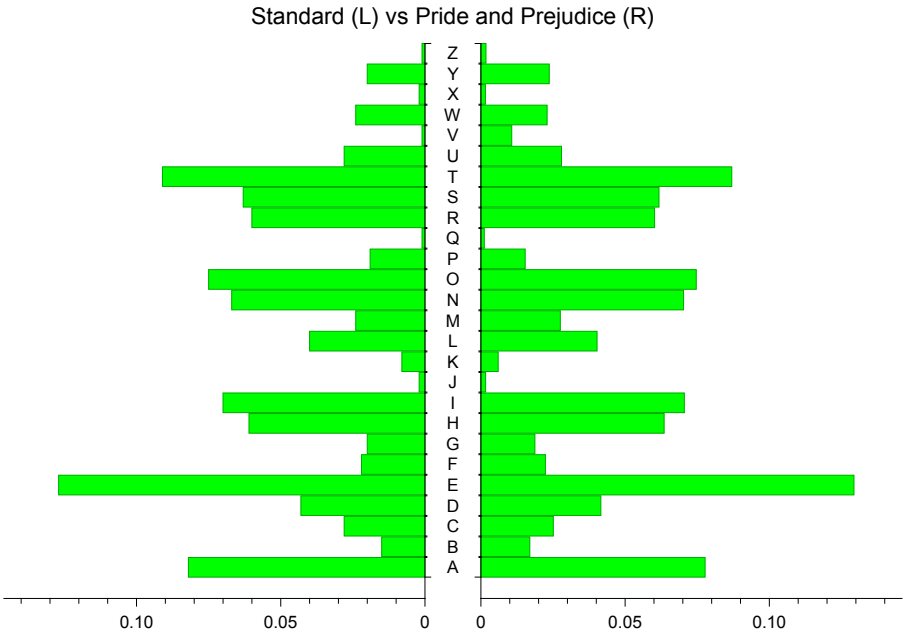

Figure 1: Distribution of letters in ’Pride and Prejudice’ compared to the standard

The second text is the Book of Genesis, taken from the King James Version, with 151 837 letters

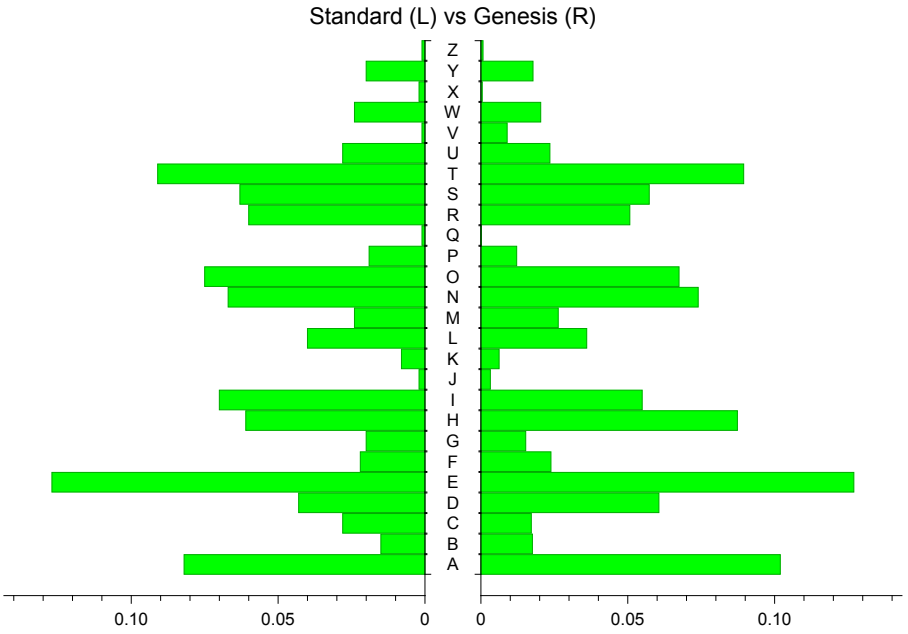

Figure 2: Distribution of letters in the Book of Genesis compared to the standard

Copared to the standard, Genesis uses '*a*' and '*h*' more frequently but '*c*' and '*i*' less frequently. The relative use of '*n*' and '*o*' is reversed. Again there is more frequent use of '*v*' in Genesis.

The next text is the Magna Carta with only 21 322 letters. As this sample size is approximately one-fifth of the standard distribution any differences that occur are not as important as might have occurred with a larger sample size. The main differences occur in the greater use of '*f* ' but less use of '*p*' in the Magna Carta.

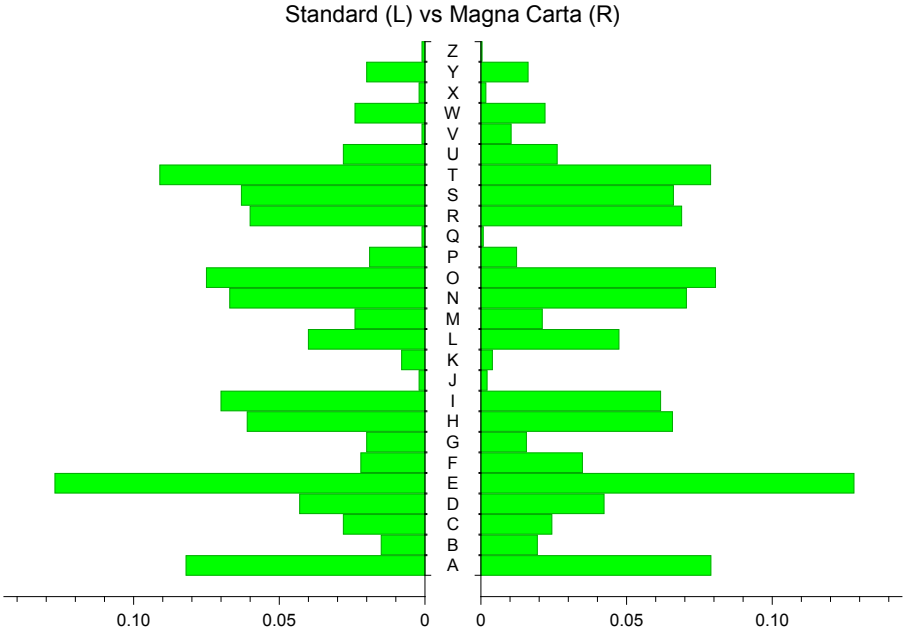

Figure 3: Distribution of letters in the Magna Carta compared to the standard

The final text is 'Origin of the Species' with 723 134 letters ie more than seven times the sample size of the standard.

Figure 4: Distribution of letters in 'Origin of the Species' compared to the standard

While there are some differences, for example the relative frequencies of '*c*' and '*d*', '*f* ' and '*g*', '*h*' and '*i*' and '*n*' and '*o*', the distributions are very similar.

Summary

Four texts from different areas of writing and with different sample sizes have shown that the distribution of the letters of the alphabet exhibits some variation but overall is consistent.

The letter '*e*' occurs most frequently, followed by '*t* ' and '*a*' then, with similar frequencies, the letters '*h*', '*i*', '*n*', '*o*', '*r*' and '*s*'. It is perhaps interesting that these are adjacent pairs. This could lead to a frequency analysis of two-letter combinations. While it is easy to think of words that include '*hi*', '*on*', '*no*', '*sr*' and '*rs*' it is not so easy to recall words that include '*ih*'. Analysis of the four texts reveals the occurrences in each, as shown in Table 2. Of interest is the relatively high number of occurrences of '*sr*' in Genesis.

String	Pand P	Genesis	MCarta	OoSpecies
hi	4488	1777	161	3383
ih	5	1	0	3
no	3350	510	74	2500
on	6021	1060	259	9126
rs	1876	371	81	1785
sr	7	44	0	5

Table 2: Occurrences of consecutive two-letter strings

Frequency analysis and ciphers

Ciphers are used to code text and thus make the meaning unintelligable. Simple ciphers that swap one letter for another can be decoded using frequency analysis, providing the text is of sufficient length.More complicated ciphers render frequency analysis useless.

One method of coding is called a stream cipher. There are various versions, for this exercise I used

an autokey cipher. Briefly, the letters have a numerical equivalent modulo 26 (‘*a*’ = 0, ‘*z*’ = 25), denoted as plaintext(*i*), and the cipher text used to code *n* letters in the plaintext is calculated from

$$\text{ciphertext}(i) = \text{plaintext}(i) + k(i) \pmod{26} \qquad i = 1, 2, \ldots n$$

where $k(1)$ is a random seed from (0, 25) and $k(i)$ = plaintext(i - 1)

Figure 5 shows the distribution of letters in ‘Pride and Prejudice’ before and after the cipher:

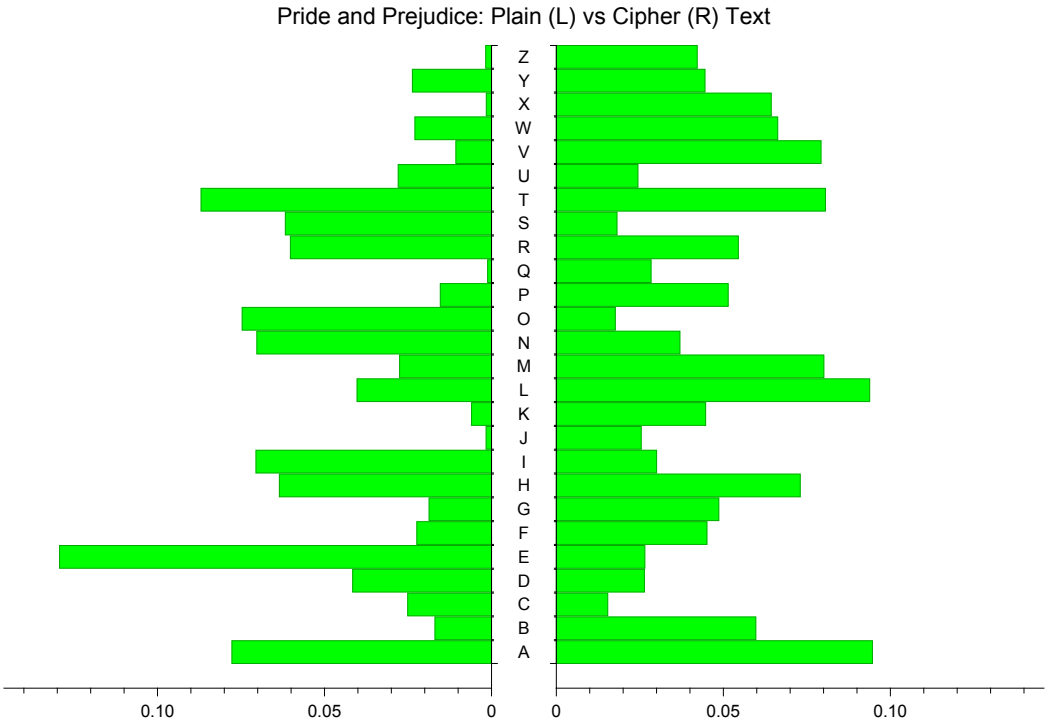

Figure 5: Distribution of letters in ’Pride and Prejudice’ compared to the cipher text

While ‘*a*’, ‘*h*’ and ‘*t*’ continue to exhibit high relative frequency in this case the remaining letters show large changes.
It is interesting to see the variation in the spread of the cipher text letters from some of the letters in the original text.
Table 3 shows the letters in the cipher text that were originally ‘*k*’, ‘*z*’ and ‘*g*’. For example, there were 3208 occurrences of ‘*k*’ in ‘Pride and Prejudice’ and this letter has been re-coded into 15 letters in the cipher text, principally ‘*d*’, ‘*k*’, ‘*m*’, ‘*y*’, ‘*x*’, ‘*v*’, ‘*s*’ and ‘*b*’

Original text	Occurrences	Cipher text
k	3208	b (112) c (98) d (670) e (1) f (2) g (13) k (585) m (428) n (1) o (88) q (2) s (236) v (238) x (359) y (375)
z	936	h (769) n (2) m (1) s (37) t (3) y (101) z (23)
g	10 031	a (862) b (14) g (843) h (1) j (122) k (322) m (41) q (2) o (874) r (18) t (4789) u (108) x (204) y (35) z (1796)

Table 3: Destinations in the cipher text of ‘*k*’, ‘*z*’ and ‘*g*’ in ‘Pride and Prejudice’.

The full changes are shown in Table 4. The letters in the original text are at the head of the columns while the letters in the cipher text are at the start of the rows. For example, ‘*k*’ in the original text was changed to ‘*b*’ 112 times in the cipher text.

	A	B	C	D	E	F	G	H	I	J
A	9	0	1	0	1789	25	862	24 130	2161	1
B	292	1537	0	171	63	3	14	0	11 136	0
C	1399	8	1121	0	352	1	0	80	357	56
D	1059	0	1	2417	33	7	0	2320	1305	0
E	3116	3	324	1	7	0	0	7	2505	3
F	776	44	0	3	4164	463	0	1	71	0
G	820	0	1632	206	2480	4	843	0	163	0
H	6301	143	0	5419	2540	0	1	8	5	0
I	708	52	18	0	1887	20	0	7	1565	0
J	299	298	0	27	1116	683	122	2252	427	0
K	17	0	1458	9	1300	572	322	3	523	18
L	1427	0	0	1318	14 823	23	0	222	1912	0
M	1879	0	0	0	1483	11	41	0	1072	1
N	550	198	29	0	234	724	0	1553	819	44
O	103	3	0	1667	1045	0	874	4	597	0
P	981	340	1651	0	3569	37	0	5	4488	6
Q	0	2	218	5584	3168	628	2	0	0	0
R	1340	0	0	538	4010	58	18	195	1	0
S	1346	54	0	0	94	224	0	0	488	0
T	14 845	55	681	0	1704	3947	4789	1	3128	0
U	451	5952	325	857	0	1	108	29	1482	0
V	512	231	4783	35	7353	0	0	99	851	0
W	2723	104	1025	3746	3809	210	0	102	335	69
X	69	0	31	254	7672	129	204	0	415	2
Y	5	0	5	41	390	4176	35	112	0	0
Z	659	63	157	6	4279	52	1796	2937	2022	0

	N	O	P	Q	R	S	T	U	V	W
A	780	1299	45	0	0	4844	920	153	0	397
B	6021	3350	706	0	0	0	4818	273	2	0
C	0	1304	21	0	28	98	0	17	0	31
D	0	1031	581	73	1129	220	0	185	880	4
E	606	0	695	4	14	295	251	0	0	1
F	19	1872	0	0	3839	1460	8	309	0	0
G	3419	1813	127	0	1337	1114	3412	717	108	15
H	1421	12 121	681	0	0	161	2684	163	0	120
I	61	7	3010	4	578	0	326	6216	250	0
J	460	318	454	251	7	1876	0	169	517	34
K	0	995	22	0	3154	1829	1288	626	0	182
L	16	0	0	0	2608	9172	3471	293	0	0
M	0	2146	292	0	7	1667	16 894	1513	212	0
N	7943	0	3	1	125	47	1833	1752	2	82
O	0	17	0	0	0	85	174	0	857	126
P	0	639	771	0	3	0	1	15	1	912
Q	61	3084	0	0	0	342	148	0	11	0
R	6161	915	0	0	4137	0	181	3	0	141
S	2	103	0	100	322	5474	0	0	0	0
T	144	2227	629	1	317	154	5650	0	0	0
U	5	570	0	187	238	7	106	509	0	3
V	10 049	2629	16	0	11 317	403	1279	1107	1363	37
W	0	2124	0	1	837	3562	0	365	1	334
X	474	152	173	0	676	3	2659	228	0	2
Y	21	2	0	5	159	269	423	21	100	0
Z	21	1323	0	0	1463	27	116	351	1423	20

Table 4: Cipher text destinations (rows) of all the letters in ‘Pride and Prejudice’ (columns)

All letters in the original text have at least one letter that is excluded from the destination letters in the cipher text. For example, ‘*a*’ is changed to every letter except ‘*q*’ in the cipher. While most letters

show a good spread of destination letters in the cipher there is usually a dominant equivalent. For example, ’*g*’ is recoded ‘*t*’ 4789 times (47.7%) with the next highest being ‘*z*’ (1796 times (17.9%)).

Of the commonly used letters perhaps the most interesting is ‘*h*’ that is recoded as ‘*a*’ 24130 times (70.8%).