Science in Context **19**(4), 591–600 (2006). Copyright © Cambridge University Press
doi:10.1017/S0269889706001074 Printed in the United Kingdom

Classical Text in Translation

An Example of Statistical Investigation of the Text *Eugene Onegin* Concerning the Connection of Samples in Chains

A. A. Markov

(Lecture at the physical-mathematical faculty, Royal Academy of Sciences, St. Petersburg, 23 January 1913)[1]

This study investigates a text excerpt containing 20,000 Russian letters of the alphabet, excluding **Ь** and **Ъ**,[2] from Pushkin's novel *Eugene Onegin* – the entire first chapter and sixteen stanzas of the second.

This sequence provides us with 20,000 connected trials, which are either a vowel or a consonant.

Accordingly, we assume the existence of an unknown constant probability p that the observed letter is a vowel. We determine the approximate value of p by observation, by counting all the vowels and consonants. Apart from p, we shall find – also through observation – the approximate values of two numbers p_1 and p_0, and four numbers $p_{1,1}$, $p_{1,0}$, $p_{0,1}$, and $p_{0,0}$. They represent the following probabilities: p_1 – a vowel follows another vowel; p_0 – a vowel follows a consonant; $p_{1,1}$ – a vowel follows two vowels; $p_{1,0}$ – a vowel follows a consonant that is preceded by a vowel; $p_{0,1}$ – a vowel follows a vowel that is preceded by a consonant; and, finally, $p_{0,0}$ – a vowel follows two consonants.

The indices follow the same system that I introduced in my paper "On a Case of Samples Connected in Complex Chain" [Markov 1911b]; with reference to my other paper, "Investigation of a Remarkable Case of Dependent Samples" [Markov 1907a], however, $p_0 = p_2$. We denote the opposite probabilities for consonants with q and indices that follow the same pattern.

If we seek the value of p, we first find 200 approximate values from which we can determine the arithmetic mean. To be precise, we divide the entire sequence of 20,000 letters into 200 separate sequences of 100 letters, and count how many vowels there are in each 100: we obtain 200 numbers, which, when divided by 100, yield 200 approximate values of p.

[1] Cf. Markov 1913a. Translated into German by Alexander Y. Nitussov, Lioudmila Voropai, and David Link; translated into English by Gloria Custance and David Link.

[2] In Russian, these letters are hard and soft signs, which are not pronounced independently but modify the pronunciation of the preceding letter.

When we determine the number of vowels, we wish to retain the possibility of constructing other combinations of 100 letters; we write down each hundred in a square with ten rows and ten columns maintaining the order of the letters:

$$\begin{array}{cccccccccc}
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\
11, & 12, & 13, & 14, & 15, & 16, & 17, & 18, & 19, & 20 \\
\multicolumn{10}{c}{\dotfill} \\
91, & 92, & 93, & 94, & 95, & 96, & 97, & 98, & 99, & 100.
\end{array}$$

Next, we count how many vowels there are in each column taken separately and join the numbers in pairs:

the 1st and 6th, 2nd and 7th, 3rd and 8th, 4th and 9th, 5th and 10th.

In this way, we obtain five numbers for each 100 letters, which we denote with the following symbols

$$(1,6), \quad (2,7), \quad (3,8), \quad (4,9), \quad (5,10);$$

and the following sum

$$(1,6) + (2,7) + (3,8) + (4,9) + (5,10)$$

represents the total number of vowels in this hundred.

If we combine 500 letters, we can construct five new groups of 100 letters each: the first – from the first and sixth column, the second – from the second and the seventh, and so on.

The number of vowels in these new groups of 100, obviously, is made up from the following sums

$$\Sigma(1,6), \qquad \Sigma(2,7), \qquad \Sigma(3,8), \qquad \Sigma(4,9), \qquad \Sigma(5,10),$$

which consist of the corresponding five summands.

The results of our counts are entered in forty small tables, each containing the following: in the first row – five numbers (1,6) and their sum, in the second row – five numbers (2,7) and their sum, etc., in the last row – the number of vowels in the first hundred, second hundred, etc., and finally the number of vowels in all five hundreds; to save space, reduced by 200:

6	8	11	11	13	49	16	11	9	8	7	51	14	12	7	3	6	42	5	11	10	6	10	42	10	6	6	6	7	35
12	11	7	7	5	42	4	8	9	11	10	42	5	5	11	9	11	41	12	8	8	11	7	46	9	12	15	6	9	51
6	6	6	7	13	38	9	9	9	7	10	44	8	10	6	10	7	41	7	7	12	10	9	45	9	3	6	10	9	37
8	10	11	9	4	42	12	9	6	10	7	44	11	11	8	3	10	43	8	12	7	9	9	45	9	11	8	5	6	39
10	11	5	10	8	44	3	8	10	8	9	38	4	4	11	14	8	41	12	8	10	9	8	47	9	10	10	10	9	48
42	46	40	44	43	15	44	45	43	44	43	19	42	42	43	39	42	8	44	46	47	45	43	25	46	42	45	37	40	10

8	7	8	7	10	40	11	11	8	7	7	44	11	10	10	12	6	49	12	9	8	10	10	49	8	9	9	5	8	39
10	9	9	8	8	44	9	6	10	11	11	47	4	4	9	7	9	33	3	10	12	9	10	44	7	9	9	11	7	43
8	9	8	8	8	41	12	9	9	5	6	41	11	13	6	9	10	49	11	11	6	11	10	49	10	6	6	9	9	40
10	6	13	6	12	47	10	8	6	11	11	46	6	7	11	8	6	38	10	8	11	6	7	42	7	8	15	6	9	45
8	12	5	13	6	44	7	6	8	9	8	38	8	6	10	7	12	43	6	8	7	9	6	36	11	7	6	11	10	45
44	43	43	42	44	16	49	40	41	43	43	16	40	40	46	43	43	12	42	46	44	45	43	20	43	39	45	42	43	12

7	7	7	7	9	37	12	7	7	6	8	40	7	4	11	5	7	34	5	5	7	5	9	31	8	6	5	14	11	44
9	13	6	8	4	40	6	8	7	10	8	39	11	14	9	11	9	54	12	6	10	10	8	46	8	12	10	7	4	41
9	7	11	12	14	53	9	10	10	8	7	44	7	6	9	8	9	39	8	14	11	11	10	54	8	10	9	8	14	49
7	11	8	9	7	42	9	5	6	7	7	34	10	9	8	10	5	42	4	3	9	5	9	30	9	5	9	9	6	38
8	10	10	11	9	48	7	11	9	13	7	47	11	10	8	9	11	49	13	14	9	11	7	54	8	13	11	5	10	47
40	48	42	47	43	20	43	41	39	44	37	4	46	43	45	43	41	18	42	42	46	42	43	15	41	46	44	43	45	19

10	9	13	6	12	50	4	11	10	12	5	42	5	11	10	6	5	37	4	4	10	11	5	34	13	11	13	10	10	57
8	8	8	9	5	38	14	9	8	7	14	52	8	9	8	10	10	45	6	12	9	8	10	45	7	10	9	6	2	34
10	10	8	9	10	47	4	8	9	8	4	33	8	8	6	9	9	40	13	4	10	8	6	41	8	8	7	8	12	43
7	9	10	7	10	43	8	14	11	12	6	51	10	6	9	7	6	38	7	10	7	12	11	47	9	11	9	10	6	45
9	8	3	11	7	38	11	6	7	4	14	42	11	9	8	10	12	50	9	13	8	1	8	39	6	3	7	9	9	34
44	44	42	42	44	16	41	48	45	43	43	20	42	43	41	42	42	10	39	43	44	40	40	6	43	43	45	43	39	13

11	6	8	9	5	39	10	10	4	7	9	40	10	8	7	8	8	41	10	3	11	13	5	42	8	8	13	5	8	42
6	10	6	8	13	43	11	10	13	13	9	56	6	9	9	8	7	39	7	11	9	7	10	44	9	10	7	14	9	49
10	5	11	11	6	43	10	7	5	9	6	37	15	9	11	13	9	57	10	10	4	7	7	38	9	11	6	8	7	41
9	12	6	8	10	45	10	5	8	10	10	43	5	10	5	4	7	31	7	7	14	13	7	48	7	9	12	6	9	43
7	11	9	10	10	47	6	13	10	5	6	40	8	9	10	12	9	48	11	9	9	6	15	50	10	9	9	12	9	49
43	44	40	46	44	17	47	45	40	44	40	16	44	45	42	45	40	16	45	40	47	46	44	22	43	47	47	45	42	24

12	7	12	5	12	48	10	14	7	6	6	43	9	6	7	10	5	37	12	13	5	9	11	50	5	11	8	12	10	46
10	8	5	13	4	40	4	6	8	10	14	42	11	10	7	8	9	45	7	7	10	5	8	37	12	8	9	8	6	43
10	13	8	7	9	47	13	6	12	8	5	44	10	10	9	9	10	48	7	7	9	14	7	44	8	11	9	8	7	43
9	4	12	6	9	40	7	13	5	8	10	43	8	6	12	10	10	46	12	13	7	8	10	50	8	5	7	11	8	39
4	12	9	9	8	42	8	5	15	10	9	47	9	11	8	5	11	44	4	4	12	11	9	40	11	11	10	6	8	46
45	44	46	40	42	17	42	44	47	42	44	19	47	43	43	42	45	20	42	44	43	47	45	21	44	46	43	45	39	17

9	11	10	6	13	49	5	9	7	10	6	37	8	6	8	7	14	43	7	9	8	6	7	37	9	11	11	8	8	47
9	8	6	8	6	37	10	9	11	7	7	44	8	14	13	8	4	47	9	8	6	10	11	44	10	8	5	9	10	42
7	7	12	10	9	45	11	11	11	10	8	51	12	4	6	9	11	42	10	9	10	8	10	47	6	8	16	12	11	53
12	12	6	8	8	46	7	7	5	10	10	39	6	8	9	10	8	41	8	7	4	9	4	32	12	11	5	7	8	43
5	7	9	11	4	36	13	8	9	8	10	48	6	8	11	8	6	39	11	8	10	8	9	46	6	5	9	10	8	38
42	45	43	43	40	13	46	44	43	45	41	19	40	40	47	42	43	12	45	41	38	41	41	6	43	43	46	46	45	23

5	7	4	3	7	26	4	7	9	11	10	41	10	8	7	8	7	40	12	10	11	4	5	42	12	13	6	6	10	47
14	10	13	9	5	51	10	7	9	4	9	39	10	8	11	10	7	46	5	9	10	11	11	46	6	3	10	10	4	33
7	8	6	8	9	38	8	13	9	12	10	52	6	11	11	10	10	48	10	8	10	7	13	48	11	11	9	7	14	52
7	10	9	5	9	40	7	5	7	7	12	38	12	8	7	6	5	38	11	8	8	11	5	43	5	8	8	9	9	39
9	10	11	16	7	53	13	10	10	9	5	47	5	9	11	12	11	48	4	8	8	9	11	40	11	6	11	12	7	47
42	45	43	41	37	8	42	42	44	43	46	17	43	44	47	46	40	20	42	43	47	42	45	19	45	41	44	44	44	18

First, we shall look at the group of numbers

$$42, \quad 46, \quad 40, \quad 44, \quad 43, \quad 44, \quad 45, \quad 43, \ldots$$

which are found in the last rows of our 40 small tables and show the number of vowels in consecutive groups of 100 letters of the text, for example:

1) **мой дядя самых честных правил когда не в шутку занемог он уважат себя заставил и лучше выдумат не мог его примѣр другим на (42 vowels)**
2) **ука но боже мой какая скука с болным сидѣт и ден и ноч не отходя ни шагу проч какое низкое коварство полуживаго забавлят ем (46 vowels)**

etc.[3]

We start a new table by counting how often each of the numbers occurs in this group.

37	38	39	40	41	42	43	44	45	46	47	48	49
3	1	6	18	12	31	43	29	25	17	12	2	1

In the first row are all the numbers that occur in the group, and in the second row beneath it, how often they appear.

With the aid of this table, the arithmetic mean is easy to find

$$43 + \frac{29 + 25 \times 2 + 17 \times 3 + 12 \times 4 + 2 \times 5 + 6 - 31 - 12 \times 2 - 18 \times 3 - 6 \times 4 - 5 - 3 \times 6}{200} = 43.19$$

and from this it follows that

$$p \neq 0.4319 \neq 0.432.$$

Next, we calculate the sum of the squares of their deviations from 43.2; it is

$$1022.8,$$

divided by 200, we get

$$5.114,$$

and this number can be regarded as the approximate quantity of the mathematical expectation of the square of the deviation of each of our 200 numbers from their common mathematical expectation, which is 43.2. Finally, the number

$$\frac{5.114}{200} = 0.02557$$

[3] This is the beginning of Pushkin's text.

represents the approximate quantity of the mathematical expectation of the square of the error when determining 100 p with the equation

$$100\,p \neq 43.2.$$

Such a deduction is associated with the usual assumption of the method of least squares, namely, that we are dealing with independent quantities. This assumption is not less justified in this case than in many others, because the connection between the numbers is fairly weak due to the way in which they were obtained.

One can also discern a certain correspondence of our results with the well-known law of error, which is associated with the names of Gauss and Laplace; for example, the quantity called probable error in our case is

$$0.67 \cdot \sqrt{5.11} \neq 1.5$$

and accordingly, between

$$43.2 - 1.5 = 41.7 \quad \text{and} \quad 43.2 + 1.5 = 44.7$$

lie 103 numbers, that is, approximately half [of the total]: 31 times the number 42, 43 times the number 43, and 29 times the number 44.

To the independence of the quantities corresponds the fact that when we combine them in twos, fours, or fives, and calculate for these 100, 50, and 40 combinations the sums of the squares of their deviations from

$$86.4, \quad 172.8, \quad \text{and} \quad 216,$$

we obtain the numbers

$$827.6, \quad 975.2, \quad 1004,$$

which do not differ very much from the number found earlier

$$1022.8.$$

Now, if we move on from samples in hundreds to single samples, we ascertain that the number

$$\frac{5.114}{100} = 0.05114$$

differs strongly from

$$0.432 \times 0.568 = 0.245376 :$$

the coefficient of dispersion (we deviate here slightly from usual terminology, whereby we should have taken the square root of the number that we call the coefficient of dispersion) is

$$\frac{5114}{24537.6} \neq 0.208,$$

that is, approximately $\frac{1}{5}$, which is explained well by the connectedness of our samples.

To clarify this connectedness, although not entirely, it will help us to calculate the above-mentioned probabilities p_1 and p_0 approximately.

We take the entire text of 20,000 letters, count the number of sequences

vowel, vowel,

and obtain the number 1104; after dividing it by the total number of vowels in the text, we get the following approximate quantity for p_1:

$$\frac{1104}{8638} \neq 0.128.$$

In the same manner, we could find an approximate value for q_0 by counting the number of sequences

consonant, consonant

and dividing it by 11,362, then $p_0 = 1 - q_0$. However, we can also substitute the tiring direct count with the following. If we subtract 1104 from 8638, we obtain the number of consonants

7534,

which follow a vowel, and as all consonants apart from the first one must follow either a vowel or a consonant, the number of sequences

consonant, consonant

is determined by the difference

$$11{,}361 - 7534 = 3827.$$

Therefore, we get the following approximate quantity for p_0

$$\frac{7534}{11{,}361} \neq \frac{7534}{11{,}362} \neq 0.663.$$

As we can see, the probability of a letter being a vowel changes considerably depending upon which letter – vowel or consonant – precedes it. The difference $p_1 - p_0$, which we denote with the [Greek] letter δ is

$$0.128 - 0.663 = -0.535.$$

Now, if we assume that the sequence of 20,000 letters forms a simple chain, then for

$$\delta = -0.535,$$

according to "Investigation of a Remarkable Case of Dependent Samples," the number

$$\frac{1+\delta}{1-\delta} = \frac{465}{1535} \neq 0.3$$

can be regarded as the theoretical dispersion coefficient; naturally, this number does not agree exactly with the previously found

$$0.208,$$

but it is closer to it than to one, which corresponds to the case of independent samples.

If we consider the sequence as a complex chain and apply the findings of the study "On a Case of Samples Connected in Complex Chain," we can make the theoretical dispersion coefficient agree still better with the experimental one.

For this, we count the number of the combinations

vowel, vowel, vowel

and

consonant, consonant, consonant

in our sequence. According to my count, there are 115 cases of the first combination and of the second – 505. When we divide these numbers by the numbers found earlier

1104 and 3827,

we get the approximate equations

$$p_{1,1} \neq \frac{115}{1104} \neq 0.104, \quad q_{0,0} \neq \frac{505}{3827} \neq 0.132.$$

With the aim of applying the findings of the above-mentioned article to our case here, we assume that

$$p \neq 0.432, \quad q = 0.568, \quad p_1 = 0.128, \quad q_1 = 0.872, \quad p_0 = 0.663,$$
$$q_0 = 0.337, \quad p_{1,1} = 0.104, \quad q_{0,0} = 0.132$$

and from these numbers we get

$$\delta = -0.535, \quad \varepsilon = \frac{-24}{872} \neq -0.027, \quad \eta = -\frac{205}{663} \neq -0.309.$$

Next, we turn to the expression of the coefficient of dispersion

$$\frac{\{q(1-3\varepsilon)(1-\eta) + p(1-3\eta)(1-\varepsilon) - 2(1-\varepsilon)(1-\eta)\}(1-\delta) + 2(1-\varepsilon\eta)}{(1-\delta)(1-\varepsilon)(1-\eta)}$$
$$= \frac{1+\delta}{1-\delta}\left\{\frac{1+\varepsilon}{2(1-\varepsilon)} + \frac{1+\eta}{2(1-\eta)}\right\} + \frac{(q-p)(\eta-\varepsilon)}{(1-\varepsilon)(1-\eta)},$$

which corresponds to the conditions of my article and is derived there.

If we insert here the values found

$$p, \; q, \; \delta, \; \varepsilon, \; \eta$$

and calculate the result, we obtain

$$0.195$$

as coefficient of dispersion, which agrees very well with the number

$$0.208,$$

found following general rules and independent of our special assumptions, so that one can hardly demand any better agreement.

Of course, we cannot claim that our example satisfies fully all theoretical assumptions; however, on the other hand, we can scarcely believe that the agreement of the numbers we have discovered is pure coincidence; rather, that it is related to a certain correspondence of the theoretical assumptions and the conditions of the example.

Now, we shall turn to the other arrangement of the 20,000 letters in hundreds that we have made. We construct a table with the repetitions of individual numbers, like the one before.

26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57
1	0	0	0	1	2	1	3	5	1	2	9	13	12	13	11	17	16	15	10	10	16	10	10	5	5	3	3	3	0	1	2

The arithmetic mean of these 200 new numbers is the same as before

$$43.19.$$

However, the sum of the squares of their deviations from 43.2 is considerably greater than before; it is, namely,

$$5788.8.$$

Here, it is necessary to pay attention to the assumption of the independence of the quantities, which is usually connected with the method of least squares (see chapter VII of my book “Calculus of Probability”); let us recall why this assumption is necessary. It is necessary to determine the weight of the end result that is expressed by equation (21), and also to calculate the mathematical expectation W, which gives the approximate value k (see my book). However, this condition will prove to be superfluous, if first, we leave aside the question of the weight of equation (21), and second, replace ξ in expression W with the number a, which we shall assume to be equal to a_0, in that we disregard the difference $a - a_0$. Then, the equations

$$\text{M. E.}\ \frac{p'x' + p''x'' + \cdots + p^{(n)}\,x^{(n)}}{p' + p'' + \cdots + p^{(n)}} = a$$

and

$$\text{M. E.}\ \frac{p'(x'-a)^2+p''(x''-a)^2+\cdots+p^{(n)}(x^{(n)}-a)^2}{n}=k$$

form the basis of our deductions,[4] not requiring any independence of the quantities

$$x',\ x'',\ \ldots,\ x^{(n)}.$$

Based on such equations and the law of large numbers, we suggest that

$$a \neq \frac{p'a'+p''a''+\cdots+p^{(n)}\,a^{(n)}}{p'+p''+\cdots+p^{(n)}}=a_0$$

and

$$k \neq \frac{\sum p^{(i)}(a^{(i)}-a)^2}{n} \neq \frac{\sum p^{(i)}(a^{(i)}-a_0)^2}{n}.$$

Only the theorem of the weight of the end result, which is expressed by the well-known equation (22), is disregarded: the weight of the result is the same as the sum of the weights of all parts.

In the given case, each of our 200 numbers represents the sum of nearly independent quantities; however, the sums themselves are connected in groups of five so that only forty of them can be regarded as independent. We have 40 groups of 500 letters each; in no group of 100 are there letters that are adjacent in the text and this is the reason for the observed independence of the parts; on the other hand, in each group the letters of the first hundred are next to those of the second hundred, those of the second hundred are next to both those of the first and those of the third, etc., and for this reason, as mentioned above, our numbers are connected in groups of five.

Under these conditions and according to the given explanations, the number

$$\frac{5788.8}{200}=28.944$$

can be considered as the approximate value of the mathematical expectation of the square of the deviations of our 200 new numbers

$$49,\quad 42,\quad 38,\quad 42,\quad 44,\quad \ldots$$

from their mathematical expectation, which is approximately

$$43.2.$$

If we pass over from the letters (samples) in hundreds to the single letters, we ascertain that the number

$$0.28944$$

[4] M. E. = mathematical expectation.

does not differ significantly from

$$0.432 \times 0.568 = 0.245376:$$

the dispersion coefficient is

$$\frac{28944}{24537.6} \neq 1.18.$$

If we now turn to the end result

$$43.19,$$

then because of the connectedness of the numbers

$$49, \quad 42, \quad 38, \quad 42, \quad 44, \quad \ldots$$

the mathematical expectation of its square of error can no longer be expressed by

$$\frac{28.944}{200} = 0.14472;$$

on the contrary, corresponding to the results of the initial arrangement of the letters in hundreds, it can be expressed (of course, approximately) by the number

$$\frac{5.114}{200} = 0.02557.$$

The connectedness of the numbers as mentioned appears when their sums are combined in twos, fours, and particularly in fives. If we calculate for these 100, 50, and 40 combinations the sums of the squares of their deviations from

$$86.4, \quad 172.8, \quad \text{and} \quad 216,$$

instead of

$$5788.8$$

we obtain the numbers

$$3551.6, \quad 3089.2, \quad 1004,$$

the last of which is nearly six times smaller than 5788.8.