Science in Context **19**(4), 601–604 (2006). Copyright © Cambridge University Press
doi:10.1017/S0269889706001086 Printed in the United Kingdom

Classical Text in Translation

On a Remarkable Case of Samples Connected in a Chain. Appendix on the statistical investigation of a text by Aksakov.[1]

A. A. Markov

I have conducted a similar investigation on a text by a different author (S. T. Aksakov: "Childhood Years of Bagrov's Grandson"). The results of this investigation, which was performed on a text passage of 100,000 letters,[2] are presented in the following tables from which one can see how and to what extent the limit theorems of the calculus of probability actually become evident.

The distribution of thousands of letters (hundreds of groups of ten) according to the groups of ten, which contain the same number of vowels.

The number of vowels in the first group of ten is shown in the first column and the number of the groups of ten in the first row. The tables provide the corresponding numbers for hundreds of groups of ten. From this, one can examine the probability that the number of vowels in the group of ten corresponds to the numbers 2, 3, 4, 5, 6, and 7 (other numbers were not found). These probabilities are entered in the penultimate column; the last column shows the values of their dispersion coefficients.[3]

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	Prob.	D.c.
2	84	15	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.0017	-
3	0	0	1	5	6	7	9	11	10	15	12	12	5	3	1	1	1	1	0	0.0835	1.19
6	0	0	0	3	6	8	5	20	12	18	10	9	2	2	3	0	0	1	1	0.0827	1.04
7	73	20	7	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.0034	-

[1] The text was originally published in Markov 1924, 577–581. Translated into German by Christiane Büchner, Lioudmila Voropai, and David Link; translated into English by Gloria Custance and David Link.

[2] [Footnote by Markov]: For the investigation I used my own handwritten copy of the text, which differs slightly from the original because of some mistakes I made. However, as these mistakes are only very minor, they should not gravely affect the results. In my first investigation, I spent much time and effort on excluding such errors. The calculations were done in both cases with the same exactitude.

[3] Prob. = probability; D. c. = dispersion coefficient.

	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	Prob.	D.c.
4	0	2	3	3	2	6	6	10	8	2	5	7	7	4	9	5	8	3	5	3	1	0	0	1	0.4276	1.02
5	2	5	6	5	3	10	8	6	8	7	11	7	6	6	0	6	1	1	0	0	2	0	0	0	0.4011	1.82

I have not put down the dispersion coefficients for 2 and 7 (they are very easy to calculate), because they do not say anything about such rare events.

The distribution of the groups of ten according to the number of vowels they contain.[4]

2	3	4	5	6	7	Prob. vow.	D.c.
17	835	4276	4011	827	34	0.44898	0.25

This distribution follows from the preceding tables and gives us the mean probability of vowels or the number of vowels in 100,000 letters and the corresponding dispersion coefficient.

The (number of) sequences, which consist of two vowels, according to my reckoning is 6588; thus

$$p_1 \neq \frac{6588}{44898} \neq 0.147,\ p_2 \neq \frac{38310}{55102} \neq 0.695,\ \delta = -0.548,\ \frac{1+\delta}{1-\delta} \neq 0.29.$$

The modified and the theoretical (below) distribution of the groups of ten according to the number of vowels.

0	1	2	3	4	5	6	7	8	9	10	Prob. vow.	D.c.
26	233	793	1699	2320	2319	1548	740	261	59	2	0.44898	1.05
26	210	771	1675	2389	2335	1586	738	226	41	3		

The modification to the order of the letters was done in the same way as in the first investigation (but without the formation of new groups of hundred): in the new groups of ten, groups of letters are combined that are separated by nine letters in the text.

[4] Prob. vow. = probability of a vowel.

The theoretical distribution of the groups of ten, which results from equation (4), refers to independent samples where

$$p = 0.44898,\ q = 0.55102,\ n = 10,$$

naturally, by applying a multiplicator of 10,000.[5]

The distribution of thousands of letters according to the number of vowels.

The first row gives the deviation (first the negative, then the positive) of the number of vowels from 449, and the second, accordingly, the number of groups of 1000 letters.

-19	17	16	15	13	11	10	9	8	7	6	5	4	3	2	1	0	+1	2	3	4	5	6	7	8	9	10	11	12	13	16	18	D.c.
1	1	1	1	1	2	1	3	3	5	5	4	5	4	7	3	7	7	4	6	3	5	0	2	2	3	5	3	1	3	1	1	0.225

The distribution of groups of hundreds according to the number of vowels.

37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52
1	2	5	21	33	69	123	163	196	171	109	50	35	10	10	2

The mean value $c^{(2)}$, $c^{(3)}$, $c^{(4)}$, $c^{(5)}$, $c^{(6)}$ *of the degree of deviation from the mean number of vowels in the group of hundreds; the dispersion coefficient and other relations.*

$c^{(2)}$ c	$c^{(3)}$	$c^{(4)}$	$c^{(5)}$	$c^{(6)}$	D.c.	$c^{(4)} : c^2$	$c^{(6)} : c^3$
4.986	0.230	83.39	11.29	2291	0.202	3.35	18.4

This table results from the numbers of the preceding table, whereby first, the deviations from 45 were taken, and then a corresponding correction was done.

The distribution of groups of hundreds according to the number of vowels, for the calculation where every second letter is left out.

In this calculation, letters that stand adjacent in the text are placed in different groups of 100, and letters that are separated by one letter in the text are put together.

[5] In the German edition of his "Calculus of Probability" (Markoff 1912), the equation used is on p. 27. Markov refers to Newton's binomial formula.

26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45
1	0	2	1	2	6	6	8	8	19	22	25	45	42	65	47	48	61	71	67

46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66
62	61	59	38	53	38	42	28	24	18	5	7	3	8	1	3	1	1	1	0	1

The mean value $c^{(2)}$, $c^{(3)}$, and $c^{(4)}$, the dispersion coefficient, and the relation $c^{(4)} : c^2$ *for the last calculation.*

$c^{(2)}=c$	$c^{(3)}$	$c^{(4)}$	D.c.	$c^{(4)} : c$
35.896	17.47	3833.5	1.45	2.97

Here, the dispersion coefficient was considerably higher than one. This fact, although not clearly stated, corresponds to the theoretical assumption about the simple chain, which was the reason for performing the last calculation.

The probabilities p_1 and p_2 for this new distribution of the letters, I calculated to begin with only for the first 10,000 letters. On average, there were 4462 vowels. The sequence of two vowels separated by one letter in the text occurred 2470 times. Thus, when we do the calculation for the letter after next, we get

$$p_1 \neq \frac{2470}{4462} \neq 0.55,\ p_2 \neq \frac{1992}{5538} \neq 0.36,\ \delta = +0.19,\ \frac{1+\delta}{1-\delta} \neq 1.5.$$

Then I did a calculation of such sequences for the entire text. This resulted in 24,773 occurrences; thus, we find that

$$p_1 \neq \frac{24773}{44898} \neq 0.552,\ p_2 \neq \frac{20125}{55102} \neq 0.365,\ \delta = 0.187,\ \frac{1+\delta}{1-\delta} \neq 1.46.$$