Application of probability and statistics methods in arrangement of railway transportation

Grigory Gefan[1,*]

[1]Irkutsk State University of Railway Engineering, 634074 Irkutsk, Russia

Abstract. Complex economic and mathematical methods are becoming more widespread in training of specialists in the field of railroad communication arrangement. The purpose of this study is to develop an effective methodology for mathematical training of railway transportation specialists on the basis of active training methods. The article deals with application of probabilistic and statistical methods to problems in design of railway transportation, for example, fluctuations in loading of railway stations and distribution of the time interval between arrival of trains. Using the example of the flow of arriving trains, the technology of testing the hypothesis that the time between arrival of trains is distributed according to the exponential law and the hypothesis of independence of events in the flow is displayed in detail. When confirming each of these hypotheses, it must be concluded that the flow of trains arriving at the station is according to the simplest (Poisson's) model. This conclusion allows using the apparatus of Markov chains to describe a random process.

Introduction

Mathematical and probabilistic models are used in various tasks related to functioning of transport (for example, [1-9]). The role of probabilistic methods in designing and analyzing the operation of railway communication is determined primarily by the fact that these methods are an effective tool for minimizing costs and ensuring traffic continuity. This is associated with the use of the traffic density concept, i.e., the number of cargoes and/or passengers per 1 km of the operated railway line or section per unit of time [10].

Average values are used when describing the density of traffic on the line or in the transport node: average number of transportation units, average time interval, average distance, etc. These values do not describe fluctuations in the traffic density. The irregularity factor is introduced to eliminate this disadvantage: it is the ratio of the largest (peak) value of the traffic density to the mean value. Determination of the irregularity factor depends on additional assumptions related to the nature of the oscillations in the traffic density. Exactly this is provided by application of probabilistic methods.

Application of probabilistic and statistical methods to problems arising in the design of railway transport is considered in this article. For example, fluctuations in loading of

* Corresponding author: grigef@rambler.ru

railway stations, distribution of the time interval between arrival of trains are among the topics considered. The first works in the field of applying the probability theory to solve the problems of railway traffic appeared before the Second World War, and monographs [11, 12] covering many issues of the mathematical theory of transport flows were published in 1960's. Then, books of German and Polish authors [13, 10] appeared in translations into Russian.

Application of probabilistic and statistical methods in arrangement of transportation is considered from the standpoint of training railway transportation specialists in this article. One can improve the efficiency of training of specialists with the following: (1) when studying the probability theory and mathematical statistics, considering specific railway tasks; (2) by applying active training methods. This determines the novelty and the main study objective - to develop the efficient methodology for mathematical training of railway transportation specialists on the basis of active training methods (business games).

It is considered mandatory that the use of active training methods is consistent with the nature of the future professional tasks and the functions of trainee, including the nature of official and job relations (the so-called contextual training theory of A. A. Verbitsky [14]). In particular, when organizing business games, there should be a "simulation of the process of activity of managers and specialists of enterprises and organizations in developing managerial decisions" [15]. It is much more difficult to ensure this in teaching mathematics or physics than when teaching professional technical or economic disciplines.

As compared to the number of attempts to use business games in such areas as economics, management, pedagogy, psychology, engineering disciplines, military science, ecology, medicine, the use of business games in mathematics is scarce. There is experience of arranging business games in the simplest sections of probability theory for students of humanitarian specialties [16], on optimization of problems for future economists [17].

A business game is an imitative gaming method of active training. In this regard, a business game developer has two main tasks. First, they must create an imitation model that would reflect the actual reality that is relevant to the professional activities of the specialist to a certain extent. Secondly, a game model must be created, which is a description of the work of participants with the simulation model. Thus, both objective and social context of the specialist's professional activity is ensured [14]. The solution of two problems related to application of probabilistic and statistical methods in design of railway transportation is considered in this article as an example of the business game structure.

Study method. Imitation probabilistic model

Let us briefly consider the essence of the probabilistic methodology as applied to the traffic density concept. One of the most important concepts of the probability theory and the theory of random processes, related, in particular, to operation of transport is the flow of events. It can be considered as a sequence of moments of onset of these events: $t_1, \ldots, t_i, \ldots$ For example, it may be the moments when the train arrives at the station. It is commonly believed that simultaneous occurrence of events is almost impossible (in this case, the flow of events is called ordinary). The stationary flow assumption is accepted, which means that the probability of occurrence of an event in time t depends only on itself t, but does not depend on the time of beginning of this interval. Therefore, if occurrence of an event is reliable within a limited period of time τ, then the probability of occurrence of this event in time t $(0 < t < \tau)$ is equal to $p(t) = t/\tau$. If, during time τ, each of the available n trains reliably arrived at the station, the probability of the event that exactly k trains $(0 \le k \le n)$ will arrive at the station within time t is determined by the Bernoulli formula [18]:

$$p_k(t) = C_n^k (t/\tau)^k (1 - t/\tau)^{n-k} \quad . \tag{1}$$

If we consider the number of trains arriving at to the station within time t as a discrete random value K, then formula (1), with $p_k(t) \equiv P(K = k)$, will describe the binomial distribution.

However, binomial distribution of the number of events within the time interval is very inconvenient for use. It is distribution (1) with two parameters (n and t/τ). In practice, the possibility of replacing the binomial distribution by another distribution of a discrete random variable (the Poisson's distribution) is often used. As it is well known, the Poisson's distribution is the limiting case of binomial distribution as the number of tests tends to infinity and the probability of occurrence of an event in a separate test tends to zero [18]. Therefore, we must admit the following in the considered problem of the arrival of trains: $n \to \infty$; $t/\tau \to 0$; $0 < nt/\tau \equiv \lambda t < \infty$. Practical research in the field of railway transport usually meets the following requirements $n > 40$, $t/\tau < 0.1$ [10]. Value $\lambda = n/\tau$ represents the average intensity of onset of events (the average number of trains arriving per unit time), and value λt is the average number of events occurring within the time interval with the duration of t. The very Poisson's distribution has the form of [18]

$$P(X = k) \equiv p_k(t) = \frac{(\lambda t)^k e^{-\lambda t}}{k!}, \quad k = 0, 1, \ \ldots \tag{2}$$

Unlike binomial distribution, this is a distribution with one parameter λt, which makes composing of tables much easier.

Wide use of analytical methods for solving problems using queuing theory tools is possible when assuming the hypothesis of the flow of Poisson's applications. To do this, it is necessary to confirm the correspondence of the real empirical data to assume the so-called simplest flow of events, which has three properties. The first two properties (ordinariness and stationarity) have already been mentioned above. The third property is the absence of post-effect: events occur independently of each other; the onset of the event does not affect the onset of other events.

Substitution of the binomial distribution with the Poisson's distribution is the transition from the model in which n points are scattered over a segment of finite length τ to the model with points scattered on the infinite time line with constant intensity λ. The Poisson's distribution of the number of trains arriving within interval t is the distribution describing an unlimited flow of events and arising in the event of rejection of the condition of a deterministic number of events in time τ. The Poisson's flow is the simplest flow, i.e., stationary, ordinary flow without a post-effect.

Random value T, which is the time interval between onset of two events in the simplest flow, has the distribution function

$$F(t) = P(T < t) = 1 - p_0(t) = 1 - e^{-\lambda t} \quad , \quad t > 0 , \tag{3}$$

and the probability distribution density $f(t) = F'(t) = \lambda e^{-\lambda t}$, $t > 0$.

Assuming that there is a simple flow of events, one can test the hypothesis of the exponential distribution of time between events. In this case, time between events has mathematical expectation $M(T) = 1/\lambda$, (λ the flow intensity), and the probability of hitting the random variable in interval (a, b) is equal to $P(a < T < b) = e^{-\lambda a} - e^{-\lambda b}$. Therefore, the statistic assessment of parameter λ will be the value opposite to the selected

mean $\lambda^* = 1/\bar{t}$, and the theoretical frequency of hitting interval number i should be calculated by formula

$$n_i' = n\left(e^{-\lambda^* t_{i-1/2}} - e^{-\lambda^* t_{i+1/2}}\right). \quad (4)$$

After this, the hypothesis of exponential distribution is tested using the Pearson's criterion. However, this is not yet equivalent to confirming the statistical hypothesis that the flow of events is the simplest (Poisson's) one. It is also required that all time intervals between events are independent random variables. Let us describe the procedure required [10]. Let the entire range of values of random variable T (the time interval between arrival of trains) be divided into partial intervals and the frequencies of hitting these intervals are known. Let us assume the following:

(a) the moment of arrival of this train is such that it is separated from the previous train by the time interval falling into partial interval i;

(b) the moment of arrival of the next train is such that it is separated from this train by the time interval falling into partial interval j;

The probability of event (a) may be assessed through its relative frequency as $p_i^* = n_i/n$; the probability of event (b) may be assessed as $p_j^* = n_j/n$. The probability of the product of these events is assessed as $p_{ij}^* = n_{ij}/n$, where n_{ij} is the frequency of the event that for two "adjacent" observations, interval i will be followed by interval j. If events (a) and (b) are independent, then $p_{ij} = p_i p_j$. In other words, "empirical" frequencies of n_{ij} must be close to "theoretical" frequencies of $n_i n_j/n$. With that, the partial intervals should be selected so that the frequencies of hitting them are not low. It is believed [10] that each frequency should be not less than $\sqrt{n}$. Comparison of empirical and theoretical frequencies in this case can also be performed using the Pearson criterion.

Source data and results

Below is the solution of two problems related to application of probabilistic and statistical methods in design of railway transportation.

1. Trains arrive at the station in accordance with the Poisson's flow of events: on an average, within τ hours, n trains arrive (i.e., the average arrival intensity is $\lambda = n/\tau$ trains per hour), and the probabilities of arrival of k trains within 1 hour are distributed in accordance with the Poisson's law. It is required to make a decision on the need to build additional arrival and departure ways in accordance with the following criterion: the probability of arrival of 6 or more trains at the station within 1 hour should not exceed a certain critical value of p_{cr}.

For example, if for $\tau = 16$ hours, $n = 25$ trains arrive (i.e., the average arrival intensity will comprise $\lambda = 25/16$ trains per hour), then the probability of arrival of 0 to 5 trains within 1 hour in accordance with formula (2) is equal to $p_0(1) = 0.210$, $p_1(1) = 0.328$, $p_2(1) = 0.256$, $p_3(1) = 0.133$, $p_4(1) = 0.052$, $p_5(1) = 0.016$, and the probability of arrival of 6 and more trains at the station within 1 hour is equal to $P(K \geq 6) = 1 - 0.995 = 0.005$. If this value exceeds the critical probability value p_{cr}, then a decision must be made on the necessity of constructing additional arrival and departure tracks.

2. Within a day, i.e., during the time of $\tau = 1440$ minutes, n trains arrive at the station: intervals between consecutive arrival time values are known (each gap is calculated as the difference between the arrival time of the train and the arrival time of the previous train). Having input the partial intervals, the hypothesis that the time between arrival of trains is distributed according to the exponential law and the hypothesis of independence of events in the flow must be tested. When confirming each of these hypotheses, it must be concluded that the flow of trains arriving at the station is according to the simplest (Poisson's) model. This conclusion is important, as it allows using the apparatus of Markov chains to describe a random process.

Table 1. Planned moments of arrival of trains at the station ($n = 84$, $\tau = 1440$).

hours	minutes	Interval (min)	hours	minutes	Interval (min)	hours	minutes	Interval (min)	hours	minutes	Interval (min)
00	07	8	07	19	3	14	23	75	19	15	18
00	14	7	07	29	10	14	52	29	19	27	12
00	37	23	07	32	3	14	58	6	19	33	6
01	03	26	07	37	5	15	02	4	19	43	10
01	07	4	07	44	7	15	08	6	19	58	15
01	19	12	08	24	40	15	15	7	20	14	16
02	10	51	08	30	6	16	01	46	20	25	11
03	19	69	08	31	1	16	07	6	20	44	19
04	23	64	08	57	26	16	30	23	20	59	15
05	16	53	09	02	5	16	32	2	21	09	10
05	18	2	09	03	1	16	42	10	21	45	36
05	32	14	09	26	23	16	56	14	22	01	16
05	58	26	09	46	20	17	00	4	22	21	20
05	59	1	10	18	32	17	06	6	22	27	6
06	04	5	10	39	21	17	10	4	22	35	8
06	09	5	10	59	20	18	08	58	22	58	23
06	09	0	11	09	10	18	16	8	23	06	8
06	19	10	11	12	3	18	41	25	23	09	3
06	27	8	11	51	39	18	43	2	23	32	23
07	15	48	13	05	74	18	48	5	23	47	15
07	16	1	13	08	3	18	57	9	23	59	12

The information on arrival of trains at the station (per day) is shown in Table 1. We could estimate the average time interval between the arrival of trains and the average intensity of events directly by this data. However, when testing the hypothesis of the distribution type, one will have to use the grouped data. Therefore, in order for the calculations to be internally consistent, let us assess the values of $\bar{t}$ and λ after grouping the data into partial intervals. Let each interval have the duration of 4 min (0-4, 4-8, …). Let us calculate the frequencies of hitting t each interval. After this, we will get the

assessment $\bar{t} = \frac{1}{n}\sum_{i=1}^{k} t_i n_i = 16.57$, where $k = 19$ is the number of partial intervals, n_i are empirical frequencies, and medium values of intervals are assumed as variants of t_i. Estimation of intensity of events is $\lambda^* = 1/\bar{t} \approx 0.06$. Now, let us calculate theoretical frequency (4) for each interval. Calculation of the Pearson criterion $\chi^2 = \sum_{i=1}^{k} \frac{(n_i - n_i')^2}{n_i'}$ gives the value of 23.47 at critical point $\chi^2_{cr}(0.05, 17) = 27.59$ (0.05 is the level of significance of the hypothesis, 17 is the number of degrees of freedom). The hypothesis is accepted: the time between arrival of trains is distributed according to the exponential law.

To test the second hypothesis, let us set three partial intervals (0-8, 8-20, 20-76) and calculate the frequencies of hitting them according to Table 1. Then, let us determine theoretical frequencies $n_{ij}' = n_i n_j / n$ and empirical frequencies n_{ij} (Table 2).

Table 2. Empirical n_{ij} and theoretical n_{ij}' frequencies.

n_{ij} / n_{ij}'	$j = 1$	$j = 2$	$j = 3$	n_i
$i = 1$	17 / 16.30	7 / 10.13	13 / 10.57	37
$i = 2$	7 / 10.13	12 / 6.30	4 / 6.57	23
$i = 3$	13 / 10.57	4 / 6.57	7 / 6.86	24
n_j	37	23	24	$n = 84$

The observed value of the criterion in our case is $\chi^2 = 10.26$. When determining the critical point of distribution χ^2, it is necessary to determine the number of degrees of freedom. In this case, it is the number of frequencies n_{ij} independent of each other. For this, let us take 5 rigid links into consideration: the sum of frequencies in the first line; the sum of frequencies in the second line; the sum of frequencies in the first column; the sum of frequencies in the second column; total frequencies. Thus, the number of degrees of freedom is equal to $m = 9 - 5 = 4$. According to the level of significance $\alpha = 0.05$, critical point $\chi^2_{cr}(0.05, 4) = 9.49 < \chi^2$; the hypothesis of independence of events in the flow must be rejected. However, with $\alpha = 0.01$ $\chi^2_{cr}(0.01, 4) = 13.28 > \chi^2$, the hypothesis of independence of events in the flow is accepted.

Conclusion

The recommended use of the aforementioned study is primarily related to improvement of mathematical training methods for railway transportation specialists. These materials can be used to develop business games and other methods of contextual training. The preferred method is to divide students into teams. 2-3 weeks prior to commencement of the game, captains are appointed from among the most enterprising students, who select the composition of the teams. The work with methodological material for the business game [19] begins, during which the captains distribute roles between team members. At the beginning of a business game, each team receives the source data (for joint analysis of the results, all teams should have the same data). On expiration of the time allotted for the task,

each team presents the protocol on a special form. The teacher arranges comparison and analysis of the results of the teams and joint discussion of errors.

Results of pedagogical experiments [20] have shown that playing a business game using probabilistic and statistical methods significantly improves the quality of training, which is confirmed by the results of intermediate testing and final examination grades. Preparation and attitude of students towards a business game is significantly different from their attitude to traditional events (workshops, control and graphic works), which is explained by emergence of a new psychological situation. In the course of this event, students not only receive the skills of independent work with references, but are also trained in disciplined team work with distribution of roles, responsibility for results of the calculations and for the decisions made. This is an effective means of translating abstract theoretical knowledge into the activity context, which makes it possible to shorten the time required for gaining practical experience drastically.

References

1. H.M. Repolho, A.P. Antunes, R.L. Church, Transportation Science, 47, 330–343 (2013).
2. D.V. Ratobylskaya, Mathematical Machines and Systems, **18**, 162-169 (2013) (in Russian).
3. G.B. Titov, Izvestia of Emperor Alexander I St. Petersburg State Transport University, **35**, 81-86 (2013) (in Russian).
4. A.V. Gasnikov, S.L. Klenov, E.A. Nurminskiy, Ya.A. Kholodov, N.B. Shamray, *Vvedenie v matematicheskoe modelirovanie transportnyh potokov* [Introduction to mathematical simulation of transport flows] (Moscow, 2013) (in Russian).
5. O.A. Sidorov, A.N. Smerdin, A.S. Golubkov, Izvestia Transsiba [Journal of Transsib Railway Studies], **31**, 123-131 (2017) (in Russian).
6. D.N. Shevchenko, Vestnik of BelSUT: Science and Transport, **35**, 37-39 (2017) (in Russian).
7. I.A. Elovoy, E.N. Potylkin, Vestnik of BelSUT: Science and Transport, **35**, 80-85 (2017) (in Russian).
8. A.A. Mihalchenko, E.P. Gurskiy, Vestnik of BelSUT: Science and Transport, **35**, 86-90 (2017) (in Russian).
9. J. Shi, L. Yang, J. Yang, Z. Gao, Transportation Research Part B: Methodological, **110**, 26-59 (2018).
10. E. Vengerskiy, *Verojatnostnye metody v proektirovanii transporta* [Probabilistic methods in the design of transport] (Transport, Moscow, 1979) (in Russian).
11. F. Height, *Mathematical theories of traffic flow* (Academic Press Ink., New York-London, 1963).
12. W. Ashton, *The theory of road traffic flow* (Methuen, London-New York, 1966).
13. G. Pottgoff, *Uchenie o transportnyh potokah* [The doctrine on transport flows] (Transport, Moscow, 1975) (in Russian).
14. A.A. Verbickiy, *Aktivnoe obuchenie v vysshej shkole: kontekstnyj podhod* [Active learning in higher education: contextual approach] (Vysshaya shkola, Moscow, 1991) (in Russian).
15. Ya.M. Belchikov, M.M. Birshteyn, Delovye igry [Business games] (Avots, Riga, 1989) (in Russian).
16. O.N. Sakharova, Vestnik Vysshey Shkoly [High School Herald], **7**, 38-44 (2008) (in Russian).
17. M.B. Sukhanov, Izvestia: Herzen University Journal of Humanities & Science, **152**, 195-202 (2012) (in Russian).

18. E.S. Venttsel, *Teorija verojatnostej* [Probability Theory] (Vysshaya shkola, Moscow, 2001) (in Russian).
19. G.D. Gefan, *Verojatnostno-statisticheskie metody na primere zadach issledovanija raboty zheleznodorozhnogo transporta* [Probabilistic-statistical methods on an example of research problems of work in railway transportation] (IrGUPS Publ., Irkutsk, 2015) (in Russian).
20. G.D. Gefan, O.V. Kuzmin, Tomsk State Pedagogical University Bulletin, **181**, 49-56 (2017) (in Russian).