Selection of the number of digits of histograms for random characteristics of the dynamics of the transport system

Y P Manshin¹, E Yu Manshina¹
¹Don State Technical University, Rostov-on-Don, Russian Federation
E-mail: elemans@mail.ru

Abstract. The article deals with the experimental studies of the machines, characterized by many random dynamic effects, both from the working bodies, and from the microrelief of the fields where the technological process is carried out, and the roads when the car is moving. The histogram of a dynamic random process, as the main means of systematization, requires optimal selection of the number of discharges to obtain the effects of factors distribution patterns. The selection of the number of discharges for the histogram distributions of experimental data sets is the key to the theoretical compile statistics. The method is illustrated using a histogram comparing a stepped histogram of an empirical distribution and a theoretical law similar to a normal one. The methods of selecting a rational choice for the number of digits is presented, that allows with the required confidence probability, to obtain correspondence of empirical histograms to the theoretical laws of distribution of random variables, for which it is proposed to use the criteria of agreement of Student, Fisher, Cochren and others. Formulas are proposed for determining the number of digits in cases of similarity of the histograms to the theoretical laws of distribution of random Gauss (normal), Rayleigh, exponential, gamma distribution, log-normal, beta distribution and other laws. Numerical testing of the method for the dynamic support reactions of the car during its movement along the microrelief is given.

1. Introduction
In solving the problems of intensifying the processing of experimental data, as well as approximating the empirical laws of distributions of random variables by some theoretical law, the issue of paramount importance is the question of choosing the number of digits of the experimental histograms \( n \) [1, 2]. A feature of the choice of the number of digits is the possibility of obtaining a statistical error associated with a partial loss of information obtained from experimental data or a decrease in the quality of information.

It is especially important to choose the right number of digits \( n \) when used as a fit test (experimental law \( P_e \) theoretical \( P_t \)) \( \chi^2 \)- criterion. Since it is based on the verification of the consent form histograms approximating curve, even when correctly chosen as the theoretical expression (based on the nomogram Pearson or any other reasons) we can reject the hypothesis correct [3]. There are two values of the factor:

1. When inflated the value of \( n \) negative role play sharp differences \( P_e \) between neighboring digits discontinuity \( P_r \) it may be associated with a transition from a continuously changing random values to discrete and failure for this \( n \) volume of the test series. [4]
2. At too low a value \( n \) there is a risk artificially distort the shape of the histogram. Example of such factors is shown in Fig. 1.
2. Methods and materials

The aim of this article is to develop a method for the rational selection of digits histogram of experimental parameters, avoiding data loss and its loss of quality, as well as the reduction of time for processing of experimental materials.

Practically solve the problem of choice \( n \) or conducting accurate approximation can brute wondering large number of digits and then reducing it to conduct for each new \( n \) check on \( \chi^2 \)-criteria until we achieve maximum value. This is a very cumbersome way the time-consuming, especially if you want to compare several approximating expressions. [5]

Estimate \( n \) it can be different. Given a sample of the one-dimensional random variable \( x \) the volume \( N \) having \( x_{\text{min}} \) and \( x_{\text{max}} \). For this sample evaluated parameters and the type of approximating differential distribution \( P_t = P_t(x) \). It is clear that when using the criterion \( \chi^2 \) its lowest value (or highest significance) give experimental histogram coinciding with 'theoretical' constructed as follows [6].

We divide the area of the random variable values for \( n_t \) equal portions of the obtained points restore perpendiculars to the axis \( x \) and replace the upper side of each trapezoid curved segment parallel to the axis \( x \) (Fig. 2).
Figure 2. Construction of a “theoretical” histogram.

Further, calculations of the number of points falling a level histogram are conducted according to the formula:

\[ r_i = \bar{X}(\Delta x \cdot P(x))N \]

Where

\[ \Delta x = \frac{x_{\text{max}} - x_{\text{min}}}{n_r} \]
\[ x_i = x_{\text{min}} + \frac{(2i - 1)\Delta x}{2} \]

\[ i = 1, 2, ..., n_r. \]

After calculating the number of dots “theoretical” histograms \( r_i \) caught in \( i \)-discharge, number of digits \( n_r \) it should be chosen so as to fulfill the criteria \( \chi^2 \) with a predetermined significance \( \lambda \).

By definition [7]

\[ \chi^2(n-z-1) = N \sum_{i=1}^{n} \frac{(F(x_i) - F(x))^2}{F(x)} , \]

where

\[ F(x_i) = P(x_i)\Delta x, \]
\[ F_i(x) = \int_{x_{\text{min}}+(i-1)\Delta x}^{x_{\text{min}}+i\Delta x} P(x)dx , \]

where \( z \) the number of estimated parameters of the distribution law.

Member \( R(x_i) = F(x_i) - F_i(x) \) represents the error calculating method of the integral numerical formula rectangles on \( i \)- the interval and can be estimated by the formula:

\[ R(x_i) = \frac{\Delta x^3}{24}P''(\eta_i) , \tag{1} \]

where

\[ \eta_i \in [x_{\text{min}} + (i-1)\Delta x, x_{\text{min}} + i\Delta x]. \]

Consequently,
\[ \chi^2(n-z-1) = \frac{\Delta \sigma^2}{576} N \sum_{i=1}^{N} \left( \frac{P'(\eta)}{P'(x)} \right)^2 < \frac{\Delta \sigma^2}{576} \cdot N \cdot B \]  

where

\[ B = \left[ \max \left\{ \frac{P'(x)}{P(x)} \right\} \right] . \]

3. Results and discussing

From expressions (1) and (2) it becomes apparent evaluation procedure \( n \):

1. to conduct a preliminary study of a selected number of random numbers, the result of which is determined \( N, x_{\text{min}}, x_{\text{max}}, \beta_1, \beta_2 \) where \( \beta_1 \) - the square of the normalized index of asymmetry \( \beta_2 \) - normalized index of diversity [8];

2. with the help of the nomogram Pearson choose approximating expression \( P_T \);

3. estimate the parameters \( P_T \) and calculate \( B \);

4. give greatest significance \( \lambda_p = 0.995 \). This is necessary because \( n \) will be evaluated on \( P_T \) and significance \( \chi^2 \)-test when checking compliance \( P_e \) and \( P_T \) it will certainly be less;

5. perform a brute force different value \( n \) as long as the inequalities \( \chi^2(n-z-1) > L \).

Where

\[ L = \frac{N}{576} \left( \frac{x_{\text{max}} - x_{\text{min}}}{n} \right)^4 \cdot B . \]

The value of the parameter of proportionality \( B \) it can be calculated for a number of laws by the following formulas [9].

Normal distribution

\[ P(x) = \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}} ; \]

\(-\infty \leq x \leq \infty ; \sigma > 0 ; -\infty \leq \mu \leq \infty . \]

\[ B_\mu = \frac{1}{\sigma^2} . \]

Rayleigh distribution

\[ P(x) = \begin{cases} \left( \frac{x}{\sigma} \right)^2 \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}} ; x \geq 0 ; \sigma > 0 \\ 0 ; x < 0 \end{cases} \]

\[ B_R = \frac{9}{\sigma^2} . \]

Exponential distribution

\[ P(x) = \begin{cases} \lambda \cdot e^{-\lambda x} ; x \geq 0 ; \lambda > 0 \\ 0 ; x < 0 \end{cases} \]

\[ B_e = \lambda^4 . \]

Gamma distribution
\[
P(x) = \begin{cases} 
\frac{\lambda^x}{G(\eta)} \cdot x^{\eta-1} e^{-x}; x \geq 0; \lambda > 0; \eta > 0 \\
0; x < 0 
\end{cases}
\]

where

\[
B_x = \frac{\lambda^x}{(\eta - 2)^x}.
\]

Lognormal

\[
P(x) = \left( A(x) \exp\left[-D(\ln x - c)^2\right]\right); x \geq 0;
\]

where

\[
A = \frac{1}{\sigma \sqrt{2\pi}}; D = \frac{1}{2\sigma^2 \ln 10}; C = \mu \ln 10; \sigma > 0; \eta > 0 \leq \mu \leq \infty; B_{ln} = I^2,
\]

where

\[
l = \frac{z^2 + z - 2D}{K^2}; K = e^{-2D} + C.
\]

But \( \alpha \) defined by the equation

\[
z^2 + (1 - 2D)z - 3D = 0.
\]

Beta-distribution

\[
P(x) = \begin{cases} 
\frac{r(\eta + \gamma)}{r(\eta) r(\gamma)} \cdot K^{\eta-1}(1-K)^{\gamma-1}; 0 \leq K \leq 1 \\
0; 
\end{cases}
\]

In other cases for \( \eta > 1, \gamma > 1 \)

\[
B_{\beta} = (\eta + \gamma - 2)6l(\mu - 1)^2(\gamma - 1)^2,
\]

for \( \eta > 1, \gamma \leq 1 \)

\[
x^2(n - z - 1) \geq \frac{r(\eta + \gamma)}{r(\eta) r(\gamma)} \cdot \frac{4^{\gamma^2}}{N} n^{\gamma + \gamma}.
\]

The calculation results for a number of time series are presented in Table 1.

**Table 1.** The results of determining the number of bits of histograms for time-series describing the dynamics of the transport machine.

| Qi type, mode number | The number of sampling points | \( n_x \) | The importance of the approximation at \( n = n_x, \% \) | \( n_o \) of the proposed method | The importance of the approximation at \( n = n_o, \% \) |
|---------------------|-------------------------------|----------|---------------------------------|---------------------------------|---------------------------------|
| 1/21                | 500                           | 10       | 0                               | 8                               | 30                              |
| 2/21                | 500                           | 10       | 2.5                             | 9                               | 35                              |
| 34/21               | 500                           | 10       | 15                              | 9                               | 60                              |
| 1/12                | 500                           | 10       | 0                               | 7                               | 30                              |
| 2/12                | 500                           | 10       | 0                               | 12                              | 55                              |
| 34/12               | 500                           | 10       | 30                              | 11                              | 85                              |

Qi - vertical loads on the right bearing axle
$Q_1$ - vertical loads on the left axle support

$Q_{3,4}$ - vertical loads on the center of the steering axle

$n = 1 + \log 2n \approx 1 + 3.322 \log N$ Sturgess formula (Raybmana).

4. Consolation

Analyzing the results shown in Table 1, it can be concluded that the proposed method of determining $n$ it works quite effectively. The proposed method reduces the time for processing of the experimental data and allows enumeration without cumbersome numbers of digits and theoretical laws to obtain a rational number of digits corresponding to the selected theoretical law [10].

References

[1] Viktorova V, Stepanyants A 2016 (Moscow: LENAND) p 256
[2] Malafeev S, Kopeikin A 2016 (St. Petersburg: Deer) p 320
[3] Shishmarev V 2010 (Moscow: Academy) p 304
[4] Cherkesov G 2012 (St.Peterburg: BKhV-Peterburg) p 480
[5] Manshin Yu, Manshina E 2018 Russian Engineering Research 38(3) 157
[6] Malafeev S 2018 (Moscow: LAN) p 368
[7] Khozyaev I 2014 Monograph: Ministry of Education and Science of the Russian Federation (Rostov-on-Don: DSTU) p 235
[8] Lisunov E 2015 Textbook for universities. Special literature (Moscow: LAN) p 240
[9] Spirina M 2018 Collection of problems: textbook (Moscow: Academia) p 176
[10] Manshin Yu, Manshina E 2019 E3S Web Conferences 135 01094
    https://doi.org/10.1051/e3sconf/201913501094