COMPLEXITY Section

A COMPLETE ALGORITHM TO STUDY THE STATISTICAL SERIES ON INTERVALS

Nicolae POPOVICIU[*], Floarea BAICU[**]

1. General notations. Example

The usual notations for a random variable X, discrete or continuous are

$$X = \begin{pmatrix} x_1 & x_2 & \cdots & x_n \\ p_1 & p_2 & \cdots & p_n \end{pmatrix},\ x_i \in \mathbf{R},\ 0 \le p_i \le 1,\ \sum_{i=1}^{n} p_i = 1$$

$$X = \begin{pmatrix} x \\ f(x) \end{pmatrix},\ x \in \mathbf{R},\quad f(x) \ge 0,\ \int_{-\infty}^{\infty} f(x)\,dx = 1 \text{ (density of probability)}$$

$$M(X) = m_X = \sum_{i=1}^{n} x_i p_i;\ \ M(X) = m_X = \int_{-\infty}^{\infty} x\, f(x)\,dx \quad \text{(mean value)}$$

$$D^2(X) = \sigma_X^2 = M(X^2) - [M(X)]^2 \text{ (variance; dispersion).}$$

For a statistical series on intervals we use the following notations

Example 1 (of statistical time series reducible to intervals). The team manager of 30 workers has marked the number of pieces produced by each worker during one day. The statistical results are given in the table 1: $w_i;\ 1 = 1,30$ (workers); $c_i;\ \ 1 = 1,30$ (absolute quantities).

Table 1.

c_i	81 90 85 100 95 105 95 98 94 94 97 99 92 87 91 96 102 86 93 103 97 109 94 89 104 98
c_i	92 87 84 108

[*] Hyperion University of Bucharest.
[**] Hyperion University of Bucharest.

2. Case 1. The time series reducible to intervals. Herbert Sturgers formula

For a statistical series on intervals we use the following notations

N = the total number of items (example: workers, students, boxes etc.

c_i = absolute value (quantities, number of pieces done by the worker i; number of exams for the student i etc.).

$c_{\min} = \min c_i$; $c_{\max} = \max c_i$; $A = c_{\max} - c_{\min}$ (the amplitude).

n = the total number of internals (groups) $[L_{k-1}, L_k)$ containing the value c_i.

$L = L_k - L_{k-1}$ (all intervals have the same length).

n_k = the absolute frequency on interval $[L_{k-1}, L_k)$.

x_k = the middle point of interval $[L_{k-1}, L_k)$; f_k = the absolute frequency on interval $[L_{k-1}, L_k)$.

Our intention is **to condense the information** from table 1 into a new table or a picture.

The following steps are called the **algorithm 1.** Now we apply it to example 1.

1) Compute $c_{\min} = 81$; 2) Compute $c_{\max} = 109$; 3) Compute $A = 109 - 81 = 28$;

4) Compute the natural number n and divide the value N = the total number of workers in n groups (intervals) with the same length L.

Method 1. $n = 1 + 3.322 \lg N$, $N \le 200$ (Herbert Sturgers formula; 1926);

$n = 1 + 3.322 \lg N = 1 + 3.322 \times 1.477 = 1 + 4.907 = 5.907$; we take $n = 6$.

Method 2. $N = 2^{n-1}$; $\ln N = (n-1)\ln 2$; $n = 1 + \dfrac{\ln N}{\ln 2}$; $n = 1 + \dfrac{\ln N}{0.693}$

$n = 5.93$; we take $n = 6$.

5) Compute the length $L = \left[\dfrac{A}{n}\right] + 1$ (integer part);

$L = \left[\dfrac{A}{n}\right] + 1 = \left[\dfrac{28}{6}\right] + 1 = [4.7] + 1 = 5$; $L = 5$. Each group contains 5 workers.

6) Condense the table 1 in table 2 with the following structure: intervals $[L_{k-1}, L_k)$, the middle point x_k, absolute frequency n_k and the relative frequency $f_k = \frac{n_k}{N}$ on each interval, the probability p_k; $k = 1, n$; $k = 1, 6$.

Table 2.

k	Interval	x_k	n_k	f_k	p_k
1	$[80, 85)$	82.5	2	2/30 = 0.067	0.067; 6.70%
2	$[85, 90)$	87.5	5	5/30 = 0.167	0.167; 16.7%
3	$[90, 95)$	92.5	7	7/30 = 0.233	0.233; 23.3%
4	$[95, 100)$	97.5	9	9/30 = 0.300	0.300; 30.0%
5	$[100, 105)$	102.5	4	4/30 = 0.133	0.133;13.3%
6	$[105, 110]$	107.5	3	3/30 = 0.100	0.100; 10.0%

The statistical series from table 1 is reduced into a statistical serie on intervals $[L_{k-1}, L_k)$.

7) Construct the discrete random variable $X = \begin{pmatrix} x_k \\ f_k \end{pmatrix}$; $k = 1, n$; $k = 1, 6$

$$X = \begin{pmatrix} 82.5 & 87.5 & 92.5 & 97.5 & 102.5 & 105.5 \\ 2/30 & 5/30 & 7/30 & 9/30 & 4/30 & 3/30 \end{pmatrix}.$$

8) Compute the mean value $M(X)$, variance (dispersion) $D^2(X)$, square deviation $D(X)$. The numerical results are:

$M(X) = 95.323$ pieces, $M(X^2) = 9133.048$, $D^2(X) = 46.574$, $D(X) = 6.824$

$M(X) + D(X) = 102.147$, $M(X) - D(X) = 88.496$.

Now we can illustrate by different pictures (designs) the information contained in table 2.

9) Construct **the diagram of repartition polygon** based on absolute frequencies or relative frequencies i.e. based on the points $P_k(x_k, n_k)$ or $Q_k(x_k, f_k)$; $k = 1, n$; $k = 1, 6$.

10) Construct **the rectangle histogram** based on the intervals $[L_{k-1}, L_k)$; $k = \overline{1,n}$; $k = \overline{1,6}$ and absolute frequencies n_k.

Remark. The purpose of number *n* is to assure a smooth histogram.

11) Construct **the circle diagram** based on absolute frequencies n_k or probabilities p_k and associate the different colors for each circle sector. If we use the values n_k, then we divide the whole circle in $N = 30$ circle sectors.

12) Compute the non-centered moments of order j $M_j(X)$; $j \geq 1$ [2].

13) Compute the centered moments of order j $m_j(X)$; $j \geq 1$ [2].

14) Compute the generating function $g_X(t)$ and the characteristic function $c_X(t)$ [2].

3. Case 2. The statistical series on intervals. Median value. Module. Quantiles. Specialized algorithms

Example 1 (of statistical series on intervals with the same length). The administrator of a lodgers house has noted the age for each inhabitant and he has included the age in one interval $[a_i, b_i)$. The results are given in table 3.

Table 3.

Nr.	1	2	3	4	5	6	7	8
Interval	[0, 10}	[10, 20}	[20, 30)	[30, 40)	[40, 50)	[50, 60)	[60, 70)	[70, a]
Absolute value n_i	18	44	68	54	42	36	16	10

We denote by V the random variable of inhabitant's age from table 3.

The problem formulation. We formulate a set of questions related with the given statistical series.

a) Count the numerical values n, N, $h =$ the step length of intervals of ages.

b) Construct **the centralized table** generated by table 3.

c) Obtain the discrete random variable *X*.

d) Compute the mean value $M(X)$.

e) Compute **the main indicators of central tendency:** the estimated mean value $\overline{X}$, the median $Me = Me(V)$, the module $Mo = Mo(V)$, the quartiles (or α – quartiles) and the corresponding sum of probabilities.

The modul is the value of random variable V having the greatest probability of appearance.

f) Compute the characteristic value for inhabitants having the age less then 30 years.

Solution. a) The values n_i are the absolute frequency; $n = 8$; $N = \sum_{i=1}^{n} n_i = 288$; $f_i =$ the relative frequency, $f_i = \frac{n_i}{N}$, $i = 1, n$; $h = 10$; $x_i = \frac{1}{2}(a_i, b_i)$ (the middle of interval L_i).

b) The centralized table (table 4).

i	1	2	3	4	5	6	7	8
L_i	[0, 10}	[10, 20}	[20, 30)	[30, 40)	[40, 50)	[50, 60)	[60, 70)	[70, a]
n_i	18	44	68	54	42	36	16	10
x_i	5	15	25	35	45	55	65	75
f_i	0.0625	0.1528	0.2361	0.1875	0.1458	0.1250	0.0556	0.0347
$x_i\ n_i$	90	660	1700	1890	1890	1980	1040	750
$x_i\ f_i$	0.3125	2.2920	5.9025	6.5625	6.5610	6.8750	3.6140	2.6025
S=Cum n_i	18	62	130	184	226	262	278	288
F=Cum f_i	0.0625	0.2153	0.4514	0.6389	0.7847	0.9097	0.9653	1.0000

S = Cum n_i in the cumulative sum of n_i, $S(1) = 18$, $S(2) = 62$, $S(3) = 130$ etc.

F = Cum f_i in the cumulative sum of f_i.

c) The discrete random variable $X = \begin{pmatrix} x_i \\ f_i \end{pmatrix}$; $i = 1, n$; $i = 1, 8$

$X \sim V$; X is a statistical approximation of random variable V; both variables describe the ages of inhabitants.

d) Mean value is $M(X) = 34.812$ ages.

e1) $\overline{X} = \frac{1}{N}\sum_{i=1}^{n} x_i n_i = 34.722$ ages.

e2) Compute the median $Me = Me(V)$. We apply **the algorithm AlgoMedian.**

Step 1. Compute the place (the location) of median

$$Loc(Me) = \frac{1}{2}(1 + \sum_{i=1}^{n} n_i);\ Loc(Me) = \frac{289}{2} = 144.5.$$

Step 2. Find the position $Pos(Me)$; frame $Loc(Me)$ in the string $\mathrm{CumS}\, n_i$

$130 < 144.5 < 184$; the value 184 indicate the position $Pos(Me) = 4 = k$.

Step 3. Find the median interval $Int(Me) = L_k = L_4 = [A, B) = [30,40)$, where $A = 30, B = 40$

Step 4. Use the median formula $Me(V) = A + h\frac{Loc(Me) - S(k-1=3)}{n_k}$

$$Me(V) = 30 + 10\frac{144{,}5 - 130}{54} = 32.685;\ Me(V) = 32.685 \text{ ages.}$$

Remark. There exists a concordance between these three kinds of media value: $M(X) = 34.812$, $\overline{X} = 34.722$; $Me(V) = 32.685$. But **the best statistical meaning** has the median $Me(V)$ for statistical series on intervals.

e3) Compute the module $Mo = Mo(V)$. Method 1. We apply **the algorithm AlgoModRelFr** based on the relative frequencies.

Step 1. Use the above median interval

$$Int(Me) = L_k; L_k = L_4 = [A, B) = [30,40).$$

Step 2. Use the formula $Mo(V) = A + h\frac{|\Delta_1|}{|\Delta_1| + |\Delta_2|}$, where

$$\Delta_1 = f_k - f_{k-1};\ \Delta_2 = f_k - f_{k+1}$$

$$\Delta_1 = f_4 - f_3 = 0.1875 - 0.2361 = -0.0486$$

$$\Delta_2 = f_4 - f_5 = 0.1875 - 0.1458 = +0.0417.$$

The final result is $Mo(V) = 35.382$ ages;

$Mo(V) = 35.382 \in L_k = L_4 = [A, B) = [30,40)$. The age 35.382 has the greatest probability op appearance.

e3) Compute **again** the module $Mo = Mo(V)$. Method 2. We apply **the algorithm AlgoModAbsFr** based on the absolute frequencies.

Step 1. Use the above median interval

$$Int(Me) = L_k; L_k = L_4 = [A, B) = [30,40).$$

Step 2. Use the formula $Mo(V) = A + h\dfrac{|\Delta_1|}{|\Delta_1| + |\Delta_2|}$, where $\Delta_1 = n_k - n_{k-1}$; $\Delta_2 = n_k - n_{k+1}$. For $k = 4$ we obtain $Mo(V) = 35.385$.

Both methods give almost the same result.

e4) Compute **quartiles** with $\alpha = 4$. The α – quartiles are Q_1, Q_2, Q_3, Any α – quartile has **place (location**) and **value.** The value is taken from the values x_i, where x_i must be ordered in increasing order. In the centralized table (table 4), the values x_i are in increasing order yet.

We apply the algorithm **AlgoQuantileN**, based on the value *N*.

Compute the quartile Q_1. *Step 1.* Find the location of Q_1:

$$Loc(\alpha; Q_1) = \frac{1}{\alpha}(N+1); \;\; Loc(\alpha = 4; Q_1) = \frac{1}{\alpha}(N+1) = 72.25.$$

Step 2. Frame the location in the string $S = \text{Cum}\, n_i$: 62 < 72.25 < 130; the value 130 indicates the interval

$$Int(Q_1) = 3 = k; \;\; L_k = L_3 = [A, B) = [20,30); \;\; A = 20.$$

Step 3. Use the computation formula

$$Q_1 = A + h\frac{Loc(Q_1) - S(k-1=2)}{n_k}$$

$$Q_1 = 20 + 10\frac{72.25 - 62)}{68} = 21.507; \;\; Q_1 = 21.507 \text{ ages.}$$

Compute the quartile Q_2. *Step 1.* Find the location of Q_2:

$$Loc(\alpha; Q_2) = \frac{2}{\alpha}(N+1); \;\; Loc(\alpha = 4; Q_2) = \frac{2}{\alpha}(N+1) = 144.5.$$

Step 2. Frame the location in the string $S = \text{Cum}\, n_i$: 130 < 144.5 < 184; the value 184 indicates the interval

$$Int(Q_2) = 4 = k; \; L_k = L_4 = [A, B) = [30,40); \;\; A = 30.$$

Step 3. Use the computation formula

$$Q_2 = A + h\frac{Loc(Q_2) - S(k-1=3)}{n_k}$$

$$Q_2 = 30 + 10\frac{144.5 - 130)}{54} = 32.685;\ \ Q_2 = 32.685 \text{ ages.}$$

Compute the quartile Q_3. *Step 1.* Find the location of Q_3:

$$Loc(\alpha; Q_3) = \frac{3}{\alpha}(N+1);\ \ Loc(\alpha = 4; Q_3) = \frac{3}{\alpha}(N+1) = 216.75.$$

Step 2. Frame the location in the string S =Cum n_i : 184 < 216.75 < 226; the value 226 indicates the interval

$$Int(Q_3) = 5 = k;\ \ L_k = L_5 = [A, B) = [40,50);\ \ A = 40.$$

Step 3. Use the computation formula

$$Q_3 = A + h\frac{Loc(Q_3) - S(k-1=4)}{n_k}$$

$$Q_3 = 40 + 10\frac{216.75 - 184}{42} = 47.797;\ \ Q_3 = 47.797 \text{ ages.}$$

Compute the **sum of probabilities** (i.e. relative frequencies f_i) corresponding to each quartile Q_1, Q_2, Q_3, denoted $S(f_i; Q_j)$.

Q_j	$S(f_i; Q_j)$	
$Q_1 = 21.507$	0.0625 + 0.1528 =	0.2153
$Q_2 = 32.685$	0.2361 =	0.2361
$Q_3 = 47.797$	0.1875 + 0.1458 =	0.3333
	The remaining difference =	0.2171.

The values of the sum are in the proximity. The quartiles Q_1, Q_2, Q_3, are good.

f) Compute the characteristic value for inhabitants having the age less then 30 years. This value is the probability $p = \frac{n_1 + n_2 + n_3}{N}$;

$p = \frac{18 + 44 + 68}{288} =$ 0.4513 = 45.13%.

REFERENCES

[1] Aniela Raluca Danciu, Mihaela Gruiescu, *Statistics. Theory and applications*, Didactic and Pedagogic Publishing House, Bucharest, 2015.

[2] N. Popoviciu, *Fundamental Chapters of Probability and Mathematical Statistics*, Victor Publishing House, Bucharest, 2014.

[3] N. Popoviciu, *Lecture notes. Problems of mathematical statistics*, Hyperion University, Bucharest, 2012.

[4] N. Popoviciu, *On central tendency indicators for a statistical series on intervals. Mathematical models. Quantiles. Specific algorithms*, will appear in the Hyperion University Annals, Series Automatics-Computer Science, Bucharest 2016.

[5] N. Popoviciu, Floarea Baicu, *Numerical applications for quantiles and central tendency indicators for a statistical series on intervals of equal length*, will appear in the Hyperion University Annals, Series Automatics-Computer Science, Bucharest 2016.

[6] Wikipedia.org / Statistics.