Proceedings of *The 1st MEDINF International Conference on Medical Informatics & Engineering, "MEDINF' 2003"*, October 9 - 11, 2003, Craiova, Romania, pp. 144-145.
Published On CD, SSN: 1454-6876, vol.5, Sup. 3 (2003)

Histograms and scales program (HISTO – version 1) for EXCEL users

Liviu DRAGOMIRESCU[1], Nicolae ZOTTA[2], Dorin CIUCĂ[3]
[1)] University of Bucharest, Str. Schitu Magureanu Nr. 1, Bucharest, Romania
liviu_dragomirescu@yahoo.com
[2, 3)] "Politehnica" University, Master of Bioinformatics, [2)] nickzotta@yahoo.com,
[3)]*cdorinf@yahoo.com* , *dorin.ciuca@inthelrom.ro*

Abstract. In paper [1] the first author presented a program to assist the medical doctor or biologist in building histograms of the numerical data series. These are necessary for various biomedical or statistical considerations. The biomedical ones may refer, for instance, to the establishment of a somatic, functional or biochemical parameter's normality limits – which only makes sense after determining the respective parameter's distribution form. The statistical ones refer to the elimination of the possible outliers; the establishment of the distribution gaussianity (normality) or of a transformed shape of it, in order to be able to later operate with the powerful parametric statistics tests.

This work contains an upgraded version of the previous paper and a methodological extension. The upgrade consists in an ad-in EXCEL spreadsheet, while the extension refers to: 1) the computation of some statistics, 2) a statistic test and 3) the drawing of classification scales in keeping with the methodology proposed by the first author in [2]. The statistics are: mean, standard deviation, coefficient of variation, skewness, kurtosis, etc. The statistic test verifies normality (gaussianity) by means of the standard errors of g-statistics. Sigmatic scales and centilic scales are being drawn, which should coincide in case the data is normally distributed.

The scales are very useful in anthropometry, to define normality standards of various biomedical parameters as well as in medicine to define different classes of measurable symptoms.

1. Introduction

The numerical data series processing is often carried out wrongly or at least inadequately. For instance, many users of the statistical programs "forget" that a mean makes sense only if the data is aggregated relatively centrally, forming a so-called unimodal distribution. The mean can be calculated for any series of numbers, and the statistical programs calculate it even when it makes no sense. Therefore, in [1] I have recommended that, before calculating the mean, one should build several histograms, which help notice the existence or nonexistence of a relatively central tendency. Only in the affirmative case the mean makes sense, (with one exception: the case of small samples, extracted by randomization for the estimation of the mean in the statistical population from which they were extracted. In this case, we'll accept any distribution of the data in the sample for the calculation of the mean, on condition, however, that we dispose in advance of the information that the variables in the population distribute with a central tendency, and therefore the mean makes sense in the population.) The same things can be said, almost identically, for the median.

Another practical problem is setting the limits of biomedical normality, respectively defining sigmatical or centilical scales for the classification of the patients from the viewpoint of a certain measurable characteristic. The option between the two scales must be made according to the following rule: we build sigmatical scales only when the data distributes cvasigaussianly, i.e. unimodally, symmetrically and with medium sharpness (equal to that of Gauss' distribution. A distribution with medium sharpness is called mesukurtic.)

2. Materials and methods

To establish the type of distribution (unimodal, bimodal or multimodal) of a measurable characteristic in a phenomenon – in a «statistical population» - based on a sample, a statistical series, I have recommended the following empirical strategy:

– Calculate the number of grouping intervals, k, given by the Sturges empirical formula ($k = 1 + 3{,}322 * \log_{10} n$, where n is the number of data in the series). Build histograms with equal grouping intervals and with conecutive numbers of grouping intervals placed around k.

– Visually examine the histograms and establish the type of each distribution, if necessary, making abstraction of the small « accidents » (see fig.2): unimodal, bimodal, multimodal. More attention should be paid to those with class numbers smaller than k, when the series'volume is small, and to the others in the opposite case.

– If more histograms with similar class numbers are of the same type we can consider that that is the distribution type « behind » the data, in the « phenomenon », i.e. in the statistical population from which the data series was extracted. In the opposite case, i.e. in case when passing from one histogram to another the distribution type changes frequently, we can consider either that the examined series has two few numbers, or that these numbers are the result of measurements made with a precision that is too high for their too small number. [1]

In case you reach the conclusion that a series of numbers is unimodal, and if the data were extracted aleatorily – to be unbiassed – and are sufficient – to describe accurately enough the population of which they were extracted – then you can build clustering scales, and normality limits respectively.

In [2] the first author proposed a sigmatical scale and a centilical scale. They were conceived so as (1) to coincide for cvasigaussianly distributed data. Moreover, on a sigmatical scale, (2) to keep the limits m-2s si m+2s, because traditionally they delimit normality, and (3) to set the « Medium » class between the limits m-0,67s and m+0,67s to include 50% of the individuals as the author proposed for the centilical scale. The following two scales resulted (Table 1), in which the classes contain the limits on their left:

Table 1. Dragomirescu's sigmatical scale and centilical scale.

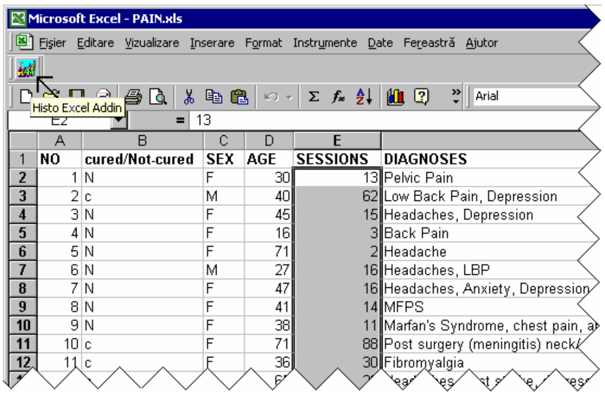
Dimension:	Very Small		Small		Medium		Big		Very Big
Sigmatical scale	→	m-2s	→	m-0,67s	→	m+0,67s	→	m+2s	→
Centilical scale	→	c_2	→	c_{25}	→	c_{75}	→	c_{98}	→

We recommend that the option between the two scales be made based on the "test for verifying normality by measuring skewness (g_1) and kurtosis (g_2)" [3]. Skewness and kurtosis are also called g-statistics [4]. The test rejects the hypothesis of normality (cvasigaussianity) if at least one of the g-statistics exceeds in absolute value 3 standard errors of the scale ($\sigma(g)$).

3. Results and discussion

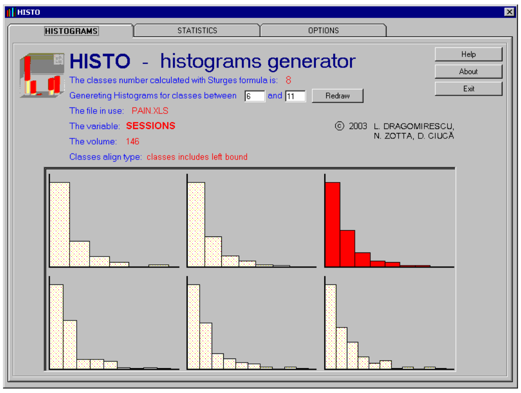
The above described biostatistical methodology was programmed in Visual Basic. The program is an ad-in for EXCEL users..

Installment: Unzip the HistoSetup.exe module and launch it. It will install in your EXCEL spread-sheet the program itself and the access pictogram (indicated by an arrow in the upper left corner of the screen in Figure 1.)


Microsoft Excel - PAIN.xls
Fişier Editare Vizualizare Inserare Format Instrumente Date Fereastră Ajutor
Histo Excel Addin
Arial
E2
= 13

	A	B	C	D	E	
1	NO	cured/Not-cured	SEX	AGE	SESSIONS	DIAGNOSES
2	1	N	F	30	13	Pelvic Pain
3	2	c	M	40	62	Low Back Pain, Depression
4	3	N	F	45	15	Headaches, Depression
5	4	N	F	16	3	Back Pain
6	5	N	F	71	2	Headache
7	6	N	M	27	16	Headaches, LBP
8	7	N	F	47	16	Headaches, Anxiety, Depression
9	8	N	F	41	14	MFPS
10	9	N	F	38	11	Marfan's Syndrome, chest pain, a
11	10	c	F	71	88	Post surgery (meningitis) neck/
12	11	c	F	36	30	Fibromyalgia

Figure 1. Histo – version 1 is an Addin program in EXCEL. The data needed for the calculation are taken by marking the area which contains them.



HISTO - histograms generator

The classes number calculated with Sturges formula is: 8

Genereting Histograms for classes between 6 and 11

The file in use: PAIN.XLS

The variable: SESSIONS

The volume: 146

Classes align type: classes includes left bound

© 2003 L. DRAGOMIRESCU, N. ZOTTA, D. CIUCĂ

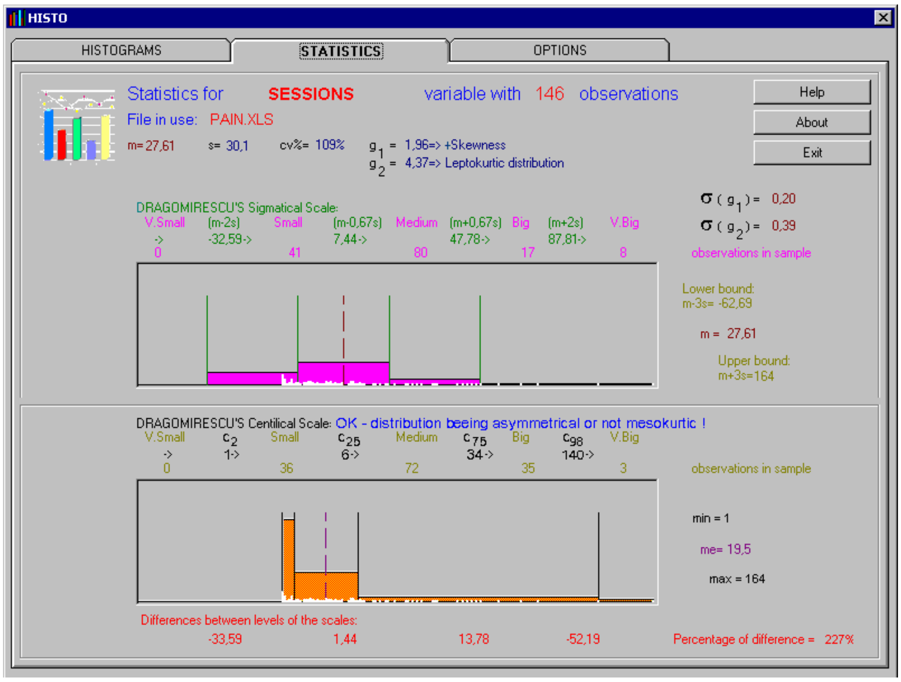
Figure 2. Unimodal distributions (with certain small « accidents » for high values), with positive skewness ⇒ the median is preferred to the mean.

Using the program is very easy:

- Mark the portion of numerical data to be analyzed and click on the pictogram. A display will appear with histograms such as that in Figure 2.

- To obtain the two scales and the related statistics click on « STATISTICS ». A display like that in Figure 3 will appear.

- Clicking on « OPTIONS » we can modify the variable's name and recalculate the classes like in EXCEL, i.e. including the limits on the classes' right.



Statistics for SESSIONS variable with 146 observations

File in use: PAIN.XLS

m= 27,61 s= 30,1 cv%= 109% g_1 = 1,96=> +Skewness
g_2 = 4,37=> Leptokurtic distribution

σ (g_1)= 0,20
σ (g_2)= 0,39

DRAGOMIRESCU'S Sigmatical Scale:

V.Small	(m-2s)	Small	(m-0,67s)	Medium	(m+0,67s)	Big	(m+2s)	V.Big	
->	-32,59->		7,44->		47,78->		87,81->		
0		41		80		17		8	observations in sample

Lower bound:
m-3s= -62,69

m = 27,61

Upper bound:
m+3s=164

DRAGOMIRESCU'S Centilical Scale: OK - distribution beeing asymmetrical or not mesokurtic !

V.Small	c_2	Small	c_{25}	Medium	c_{75}	Big	c_{98}	V.Big	
->	1->		6->		34->		140->		
0		36		72		35		3	observations in sample

min = 1

me= 19,5

max = 164

Differences between levels of the scales:

-33,59 1,44 13,78 -52,19

Percentage of difference = 227%

Figure 3. The indicated scale is centilic. The limits of the two scales differ very much, by 227%.

Figure 3 shows that the sigmatic scale for the number of sessions in neurofeedback training [5] would be an aberration: the lower limit for the « Small » class would be negative. The program recommends the centilic scale, as the distribution is "with positive skewness and not mesokurtic". It is leptokurtic, i.e. sharper than Gauss' distribution. In the last line of the display the immense difference between the two scales is obvious (227%).

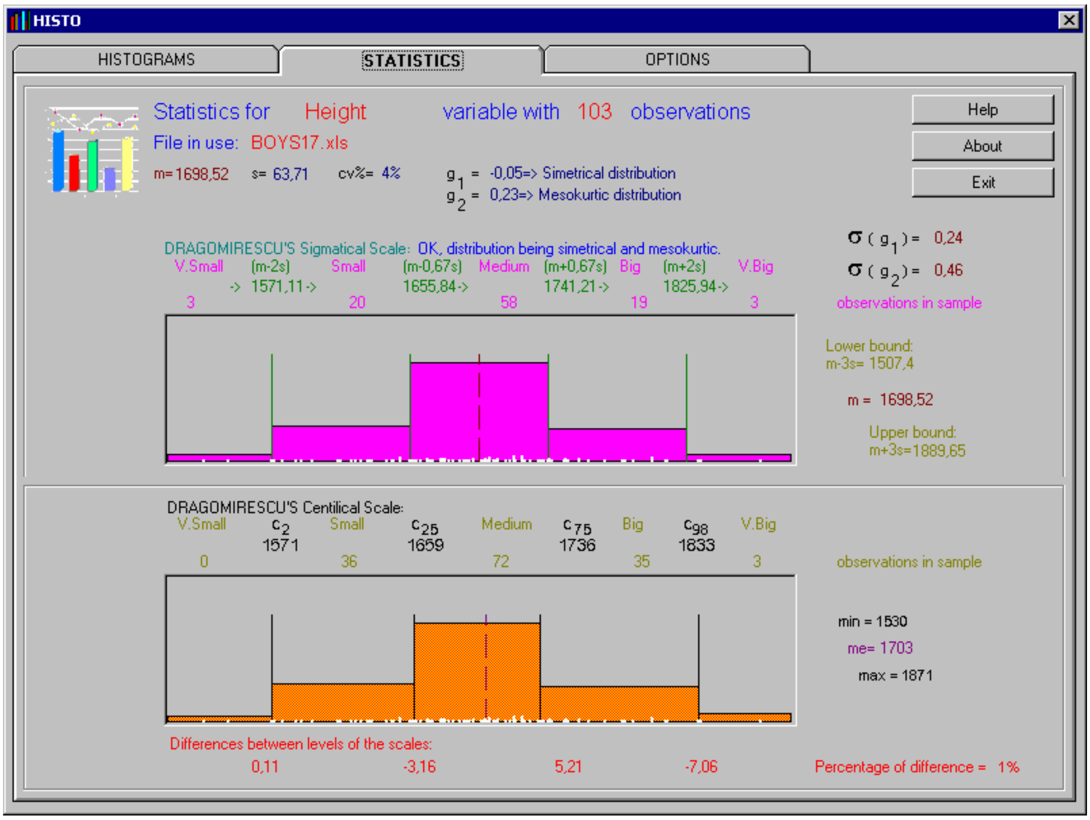
The various methodological statistical details (including the calculation of the « percentage of difference"), as well as the user instructions of the product are available in "Help".

4. Conclusion

The Histo – version 1 program is more than a calculation program. It has two additional functions:

1. It helps the user not make some of the most serious and unfortunately frequent statistical mistakes: the calculation of a measure of central tendency when there is no central tendency;
2. It has a demonstrative role, allowing the user to compare the two scales built precisely to be compatible, coinciding in case of normality.

For instance, in Figure 4 the two scales for the « Height » are built for a homogenous series of 103 boys aged 17. It is visible that the scales are almost identical, the difference being only 1% [2].



Statistics for Height variable with 103 observations

File in use: BOYS17.xls

m=1698,52 s= 63,71 cv%= 4%

g_1 = -0,05=> Simetrical distribution

g_2 = 0,23=> Mesokurtic distribution

$\sigma(g_1)$= 0,24

$\sigma(g_2)$= 0,46

DRAGOMIRESCU'S Sigmatical Scale: OK, distribution being simetrical and mesokurtic.

V.Small	(m-2s)	Small	(m-0,67s)	Medium	(m+0,67s)	Big	(m+2s)	V.Big	
	-> 1571,11 ->		1655,84 ->		1741,21 ->		1825,94 ->		
3		20		58		19		3	observations in sample

Lower bound:
m-3s= 1507,4

m = 1698,52

Upper bound:
m+3s=1889,65

DRAGOMIRESCU'S Centilical Scale:

V.Small	c_2	Small	c_{25}	Medium	c_{75}	Big	c_{98}	V.Big	
	1571		1659		1736		1833		
0		36		72		35		3	observations in sample

min = 1530
me= 1703
max = 1871

Differences between levels of the scales:

0,11 -3,16 5,21 -7,06

Percentage of difference = 1%

Figure 4. In this case one can use the sigmatic scale and the normality limits.

References

[1] L. Dragomirescu, Histogram Building Program (HISTO) According to an Original Methodology Proposed for Physicians. In: G. I. Mihalaş et al. (ed.), Healthcare Telematics Tupport in Transition Countries - Proceedings of MIE 2001 Special Topic Conference. ISBN: 973 8181 91 7. Publishing House EUROBIT Timişoara, 2002, pp. 127-132.

[2] L. Dragomirescu, Cristiana Glavce and Elena Radu, Ghid practic de antropologie. Vol. II. Prelucrarea primară a datelor antropometrice. Editura Ars Docendi, Bucureşti, 2002.

[3] M. Iosifescu et al., Mică enciclopedie de statistică. Editura Ştiinţifică şi Enciclopedică, Bucureşti, 1985.

[4] D. Nelson (eds.), The Penguin Disctionary of Mathematics. Secon Edition, Penguin Books, 1998.

[5] Victoria L. Ibric, L. Dragomirescu, Neurofeedback training in chronic pain syndromes. Workshop presentation in „The 11th Annual iSNR Conference", September 18-21, 2003, The Galleria at Houston, Texas, USA.