Use of GeoGebra in teaching about central tendency and spread variability

Pellumb Kllogjeri and Adrian Kllogjeri

ABSTRACT. In data analysis are used graphs, tables, and numerical summaries to study the variation present in data under study. Graphs are effective visual tools because they present information quickly and easily. Graphs are commonly used by print and electronic media and, data can be better understood when presented by a graph than by a table because the graph can reveal a trend or comparison. Graphs display the relationship between variables or show the value spread of a given variable or phenomenon. Measuring the central tendency we identify the most representative value in a set of data. Mean, median and mode give different perspectives of a data set's centre, but a data description is not complete until the spread variability is also known. The basic numerical description of a data set requires measures of both centre and spread. In this paper we will be concentrated on graphical representation of data collected from a population. A useful tool to get such nice visual representations is GeoGebra. GeoGebra is an interactive geometry system and, by the use of its virtual tools, we can do constructions with points, vectors, segments, lines; we can construct histogram and geometrically show the positions of the median and quartiles [4]. GeoGebra provides numerical calculations for the measures of data centre and spread. GeoGebra software is very useful and helpful for teachers and students to graphically present data and easy understand and interpret them [6]. Also, GeoGebra is dynamic mathematics software for schools that joins geometry, algebra, and calculus. Using this dynamic feature in the teaching hour of statistics, the students have a better and fuller understanding about data measures and data spread.

1. PRELIMINARY WORK

Univariate analysis involves the examination across cases of one variable at a time, specially looking at three major characteristics of a single variable that are: the distribution, the central tendency, the dispersion [1].

The Distribution is a summary of the frequency of individual values or ranges of values for a variable. A typical way to describe the distribution of the marks of the students of the experimental class is as in the table below. Distributions may also be displayed using percentages, so we describe the level in mathematics by listing the number or percentage of the students in accordance with the group or range of the points they belong to (the values are grouped into ranges and the frequencies determined). This is called a frequency distribution for grouped data. We have brought here two examples of data from the experiment done with a class of a secondary school regarding the impact of GeoGebra software in teaching and learning mathematics. To compare the two classes by using the same measurement tools we have taken into consideration the points and the marks at the beginning of the chapter in order to display their histograms back to back. We have prepared the following distribution of marks and the respective percentages for the experimental and control class (Table 1):

Range of points	Mark	Experimental Group Frequencies	Percentage Percentage	Control Group Frequencies	Percentage
< 35					
35 - < 45	4	1	3.40%		
45 - < 55	5	3	10.34	2	7.14%
55 - < 65	6	4	13.79	1	3.57%
65 - < 75	7	2	6.90	6	21.43%
75 - < 85	8	8	27.60	7	25%
85 - < 95	9	10	34.50	6	21.43%
95 +	10	1	3.40	6	21.43%
		Sum = 29		Sum = 28	

TABLE 1. Frequency and percentage distribution table.

As can be seen from the percentages in the table the control group has a better distribution of percentages and a higher level in mathematics (for higher marks, starting with the grade 7, it has higher accumulative percentage). This is the state of the groups in the beginning of the experiment. In the table above we find only figures and one cannot get a clear information and cannot do good analysis of data only by looking at them. The geometric representation is very important for the students of every cycle system of education. To master the students in calculating different statistics is the same as in the case of making sense of formal concept definitions which have to be linked with the concept images. The concept images of the students are based on their prior knowledge got through their different experiences.

"The tendency of many students to evoke their concept image[] in many situations it is desirable to have and evoke rich concept images and,.research shows that visualization facilitates mathematical understanding" (Handbook of Research on the Psychology of Mathematics Education, Advanced Mathematical Thinking, pg.149) [3].

The same frequency distribution we have depicted in a graph that is often referred to as a histogram or bar chart as shown in (Figure 1).
Using Geogebra tools we have constructed back to back histograms (called bi-histogram) for the two groups in order to make easier the reading and interpretation of data. The histogram above the horizontal axis, is of the experimental group, the histogram below the horizontal axis is of the control group. By the graphical examination done, the above-axis histogram does not appear to be that much different from the below-axis histogram. With respect to distributional shape, note that the two histograms are skewed right. Thus the bi-histogram reveals that there is not a clear difference between the two histograms with respect to location and distribution, whereas, in regard to variation there is a little change of 1 unit. To get clearer information about data we look at the other estimators which are:

Central Tendency. The central tendency of a distribution is an estimate of the "center" of a distribution of values [1]. There are three major types of estimates of central tendency: Mean, Median and Mode. By arranging the marks of the two groups in the previous chapter in increasing order is found out that, the computed **means** are: for the experimental group $\bar{X} = 7.6$, for the control group $\bar{X} = 8.14$. **Median** is the score found at the exact middle of the set of values. Using the formula of the median positioning the calculated

median is 8, which is the mark dividing into two equal parts the given sequence. For the control group the computed median is also 8. The two groups have the same median. Mode is the most frequently occurring value in the set of scores. Considering the ordered scores as shown above, and then counting each one, is found out that, the experimental group has two modes(is bimodal) : 8 and 9; the control group has one mode: 8

Summarizing: for the experimental group the mean, median and mode are 7.6, 8, 8 and 9, respectively; for the control group they are 8.14, 8, 8 respectively. From the above bi-histogram, is seen that the initial state of the experiment is: the two sets of data are centered at the value of approximately 8. That indicates that these sets are not displaced from one another, thus it is difficult to conclude that there is a real difference between the two groups.

2. USE OF GEOGEBRA IN DISPLAYING HISTOGRAM AND CALCULATION OF STATISTICS

In a simple histogram, the axis is split into sub-intervals of equal width called **classes**. The classes must satisfy the condition of non-overlapping. A rectangle is drawn above each class with height equal to the number of values in the class which is the **frequency** of the class [5]. Here, we present the case when all histogram classes are of equal width. The histograms are often drawn with a vertical axis giving the frequencies for each class. Based on data of (Table 1) and where it is clear what are the extreme values , the boundaries of the classes are: 35, 45, 55, 65, 75, 85, 95 and 105 (representing points). The classes are: 35 points to less than 45 points, 45 points to less than 55 points etc. The first class defines the mark 4, the second one defines the mark 5 and so on. The marks are the centers of the classes. To adjust the class boundaries (that are in points) with the marks we have converted them by dividing by 10. So, the classes are: 3.5 to less than 4.5, 4.5 to less than 5.5 etc. An alternative is to label the axis with the **proportions** of values in the classes. In this case each rectangle height is made equal to the **proportion** in that class instead of the class frequency. These proportions are also called **relative frequencies**. An individual relative frequency histogram has the same shape as the corresponding frequency histogram. The bar heights are simply divided by the total number of observations which rescales the histogram height. Using relative frequencies allows us to make more meaningful comparisons between the distributions of different groups.

To compare the distributions in two groups of values (e.g. measurements for the experimental group and control group), histograms for the two groups can be superimposed on the same axes. We get this way the so-called bi-histogram. Color or shading should be used to help distinguish the two histograms.
In the (Figure 1) below, the light rose histogram, above the horizontal axis, corresponds to the experimental group and the blue histogram, below the horizontal axis, corresponds to the control group. In the same picture we have displayed the box plot and the five summary numbers (two extremes, median and two quartiles).

Procedure for displaying the histogram of data of experimental group.
(1) Open e new GeoGebra document
(2) Show the coordinate axes and the grid on it
(3) In the input field apply the command histogram and register in it the Class boundaries and Heights in accordance with the rule: Histogram [List of Class Boundaries, List of Heights]. By observing this rule for entering data is created
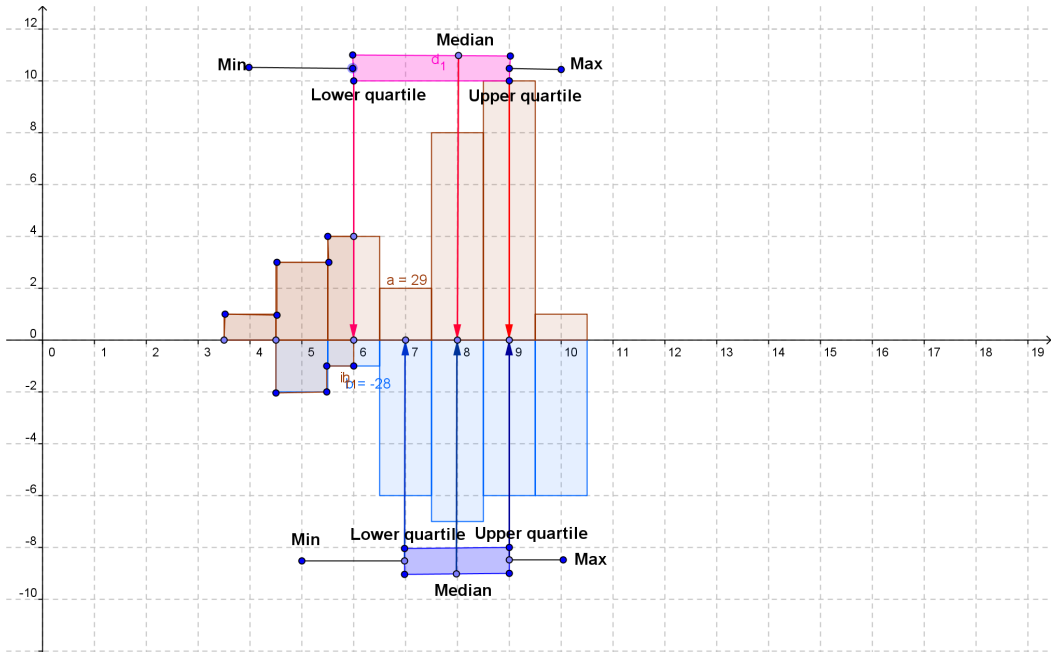
FIGURE 1. *Histograms (bi-histogram) and their respective box plots*

a histogram with bars of the given heights. The class boundaries determine the width and position of each bar of the histogram. In our example for the experimental group data we have:

Histogram $\{[3.5, 4.5, 5.5, 6.5, 7.5, 8.5, 9.5, 10.5], [1, 3, 4, 2, 8, 10, 1]\}$. After hitting the Enter key, is created the upper X-axis histogram. The first bar is positioned at the interval $[3.5, 4.5]$, the second bar is positioned at the interval $[4.5, 5.5)$, and so on. The right boundary of a class belongs to the next class.

(4) In the input field apply again the command histogram and register in it the Class boundaries and Heights for the control group (in accordance with the rule: Histogram [List of Class Boundaries, List of Heights]). The same way is displayed the histogram corresponding to control group (blue color one)

The **Box plot** of each one of the sets of above values displays these five values graphically [1]. A box plot splits the data set into four quarters with (approximately) equal numbers of values. Looking at the histogram above the horizontal axis in (Figure 1), we see that: **Centre**- the vertical line inside the box (the median) gives an indication of the centre of the distribution: the centre is 8.

Spread- the width of the box (the inter-quartile range) gives an indication of the spread of values in the distribution. It is 3. **Shape**- the high density of the values is found in the part of histogram having the median and the upper quartile. The right extreme and quartile are closer to the median than the left extreme and quartile, this shows that the distribution is right skew.

How are defined and calculated these statistics in the case of grouped data?

Our above example is far of a case where in each class of the grouped data are a great number of data. We have just 29 items of data. However they are grouped in classes and, it is not wrong to think of each class as having a great number of items, meaning that the height of the class will be greater. The procedure we will apply for defining them is the same. Consider the case of the experimental group linked with the upper X-axis histogram.

Procedure:

(1) *Follow the steps (1) - (4) of the previous procedure to construct again the histogram (green color one). *** After these steps, we construct the histogram with the same classes and for accumulated frequencies, called accumulated histogram. Further steps,*

(5) *In the input field apply the command histogram and register in it the Class boundaries and Accumulated Heights (accumulated frequencies) in accordance with the rule: Histogram [List of Class Boundaries, List of Accumulated Heights]. By observing this rule for entering data is created a histogram with bars of the accumulated heights. The class boundaries determine the width and position of each bar of the histogram. In our example for the experimental group data we have: Histogram $[\{3.5, 4.5, 5.5, 6.5, 7.5, 8.5, 9.5, 10.5\}, \{1, 4, 8, 10, 18, 28, 29\}]$. After hitting the Enter key, is created the upper X-axis accumulated histogram (light pink,(Figure 2)). Here must be understood that the class $45 - 55$ rectangle of accumulated histogram is the sum of the class $45 - 55$ rectangle of the histogram (green) with the class $35 - 45$ rectangle of the histogram (green). The class $55 - 65$ rectangle of accumulated histogram is the sum of the class $45 - 55$ rectangle of the accumulated histogram (light pink) with the class $55 - 65$ rectangle of the histogram (green) and so on. Operating this way we reach to the last rectangle of class $95 - 105$ of the accumulated histogram with area 29 square units, equal to its height.*

Having into consideration the usual practice for defining the statistics by arranging data in increasing order(from the lowest value to the highest), ordering after which are defined the positions of median and quartiles, we arrange the rectangles of the histogram (green) in a vertical way. This is achieved by interchanging the positions of the rectangles of the histogram (green) with the respective rectangles above them in the accumulated histogram (light pink). The new positions of the rectangles of the histogram are those with blue color. It is clear that these rectangles have only one vertex in common and the sum of their area is 29, equal to the area of the rightmost rectangle of the accumulated histogram. Arranging the histogram rectangles this way makes easy for us the defining of the statistics.

(6) *Step 6 is: arrange the histogram rectangles by putting them in the positions (blue) in the accumulated histogram (Figure 2).*

(7) *Construct a segment of length 29 perpendicular to X-axis, (BC). This segment is used to measure the statistics, so is not wrong to call it "**Ruler of statistics**".*

(8) *Divide it in four equal parts. The division points D, E, F correspond to upper quartile, median and lower quartile respectively. The division of the segment in equal parts is done by using GeoGebra tools and mathematical assertions.*

(9) *In points C, D, E, F, B construct the verticals p_1, k_1, l_1, m_1, n_1, respectively (they are perpendicular to segment (BC)).*

*(10) Define the intersections of these verticals with the middle height of the respective rec-
tangle (blue). The respective blue rectangle of a vertical is that one which is intersected
by the vertical. Is taken the middle height of the rectangle because it is known that
the value of a class is the middle value (point) of that class.*

*(11) From the intersection points construct the perpendiculars to X-axis. In (Figure 2) they are
represented by the arrows and just for quartiles and the median.*

*** The end points of the arrows on X-axis show the values of the quartiles and median
which are: 9(upper), 8 and 6(lower). Also, it is easy to see that the extreme values are:
10(the highest) and 4(the lowest). From the picture can be understood that the three im-
portant statistics divide the set of ordered data into four equal parts. This property can be
easily demonstrated by comparing the four parts in which the three verticals divide the
right side rectangle of the accumulated histogram (it has a height of 29 units). We can con-
struct polygons with sides the sides of each part and read the numbers corresponding to
their areas (this is another feature of GeoGebra, showing in algebra window the algebraic
representative of the geometrical object). These numbers are approximately equal.

In summary, the Experimental group five values are:
$Min = 4$, $Q1 = 6$, $Med = 8$, $Q3 = 9$, $Max = 10$

Further manipulations and observations on central statistics

With GeoGebra it is possible to create for the students new situations in order they be
able to cope with them and find the values of the statistics. Keeping the overall quantity
of data fixed (for example, $\sum f(x_i) = 28$) we can change the frequencies of the classes
and see what are the values of the statistics. We have purposely chosen a case with the
same classes as in the example above and with frequencies: $7, 0, 0, 4, 9, 0, 8$. There are
gaps between classes and it is such example where the vertical corresponding to the lower
quartile intersects the heights of two classes with center 4 and 7, respectively (Figure 3). In
this case the arrow corresponding to this quartile must have as origin the middle point of
the segment touching the rectangles of these two classes, that is, the point of X-axis with
abscissa $(4 + 7)5 = 5.5$. Further, continue by right clicking on the label of the histogram,
on the dialog box on Object Properties, then on Basic of the window open, where are
changed the class frequencies (not class intervals) and apply the changes. The same thing
is done with the label of the respective accumulated histogram. Using the ruler of statistics
(because the sum of the frequencies is fixed: 28) where the positions of statistics are fixed,
can be read the values of the new statistics and so on.

The **Standard Deviation** is a more accurate and detailed estimate of dispersion because
there are cases of ordered data with outliers that can greatly exaggerate the range. The
respective geometric positions of the mean and the one St. Deviation interval are in X-axis
and to find out the percentage of the values of the set belonging to this interval we have
to construct verticals to X-axis at the extremes of the interval and estimate the area of the
histogram included in.

ADVANTAGES

(1) GeoGebra software provides a quick construction of the histogram and a nice pic-
ture(different styles and colors)
(2) GeoGebra provides a geometrical interpretation of the statistics that cannot easy
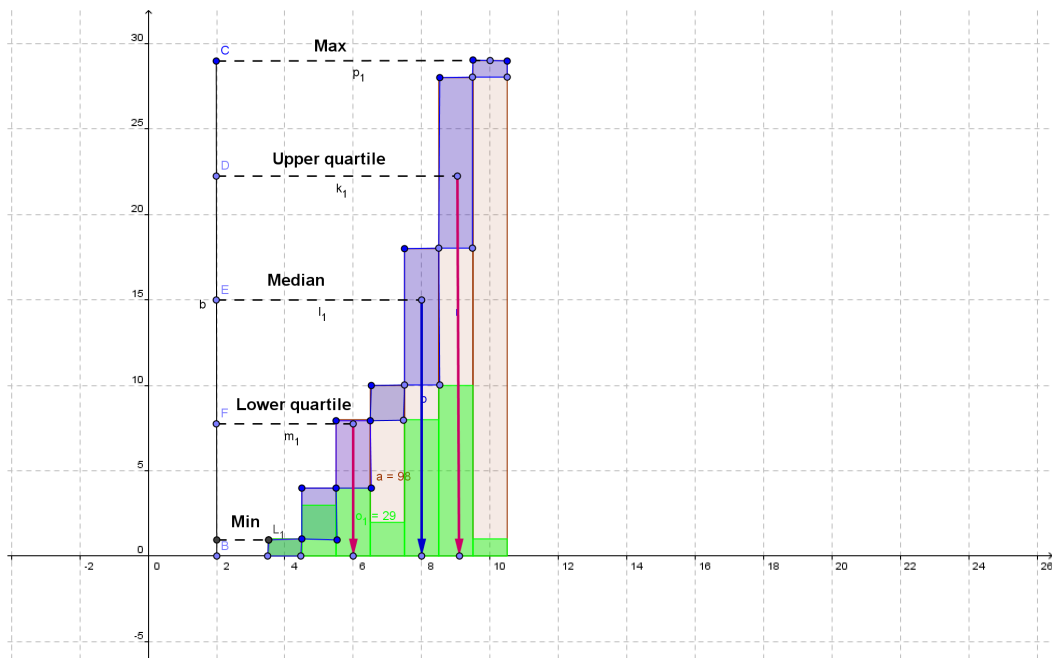be performed by other program.

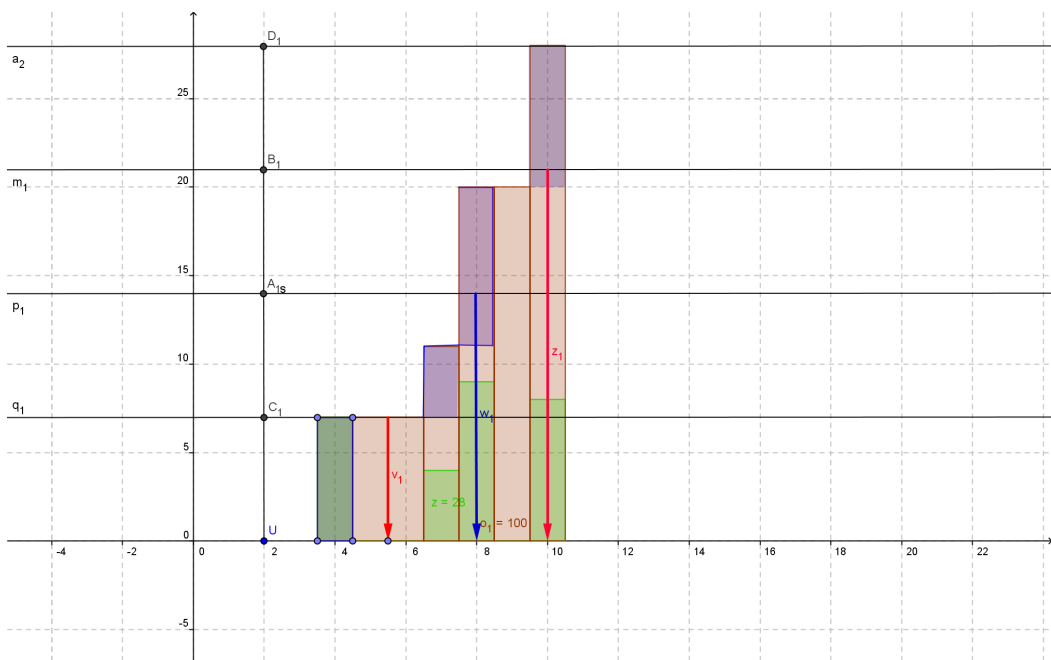FIGURE 2. *Histogram, accumulated histogram and the ruler of statistics*

FIGURE 3. *Histogram with gaps and the ruler of statistics*

(3) The use of the ruler of statistics makes easier and faster the procedure for finding the central statistics.

(4) The use of the ruler of statistics helps the students not make mistakes in reading the values of statistics in each case of the frequency distribution versus the method of using formulas for the grouped data by which can make mistakes when there are gaps between the classes

(5) Teachers and students have the possibility to dynamically demonstrate and understand the relation between the frequencies and statistics.

(6) It is possible to construct on the same pad more than two histograms in order to make comparisons between different data sets, to save the work and, the more important, to print it out for purposes of use in papers and books.

REFERENCES

[1] David, H. Voelker and Peter, Z. Orton, *Statistics*, Copyright 2001, Wiley Publishing, Inc., 11–36

[2] GeoGebra wiki sources

[3] Gutierrez, A. and Boero, P., *Handbook of Research on the Psychology of Mathematics Education* (2006), Published by: Sense Publishers, The Netherlands, 140–158

[4] Hohenwarter, J. and Hohenwarter, M., *Introduction to GeoGebra*, Copyright 2008, Pg. 16, 53, 54

[5] Lavicza, Zs. and Hohenwarter, M., *Establishing a Professional Development Network*, National Center for Excellence in the Teaching of Mathematics, 16–18

[6] Spiegel, M. R., *Statistics* 2/ed, Copyright 1988, McGraw-Hill Book Company, 58–86

DEPARTMENT OF MATHEMATICS AND INFORMATICS
UNIVERSITY ALEKSANDER XHUVANI
ELBASAN, ALBANIA
E-mail address: pkallogjeri@gmail.com
E-mail address: adrian86k@yahoo.com