HOW TO READ AND GENERATE SCIENTIFIC PUBLICATIONS. IN THIS ISSUE: GRAPHICAL EXPLORATION OF QUANTITATIVE DATA: THE IMPORTANCE OF LOOKING AT THE INFORMATION

MSc. Eng. Mauricio Fuentes A.[1], Kins, MPh. Karla Yohannessen V.[2-3]
1 Assistant Professor, Biostatistics Program, School of Public Health,, Universidad de Chile
2 Assistant Professor, Environmental Health Program, School of Public Health, Universidad de Chile
3 Assistant Professor, Department of Pediatrics and Child Surgery, School of Medicine, Universidad de Chile

ABSTRACT

Once the data collection of a study is finished and the respective database is available, the researcher is often impatient to answer the research question and ventures into the final steps of the analysis. However, a key stage, prior to a more complex or sophisticated statistical analysis, is data exploration and descriptive statistics. Unfortunately, the exploratory analysis of the data is often performed without much dedication, or just simply "skipped", which can have important consequences on the results obtained and lead to the report of erroneous conclusions. On the one hand, the exploration allows to detect errors in the data and, if possible, correct them from the source of origin or take them into account to make decisions about what to do with them. On the other hand, exploration allows to learn about the behavior of the variables evaluated in terms of their distribution (key concept in Statistics) and possible relationships between them, which is essential for subsequent descriptive and inferential analysis. The objective of this article is to show graphical tools for the exploration of quantitative data, in order to visualize its distribution and compare groups according to categories of qualitative variables.

Keywords: quantitative variable, exploratory analysis, statistical graphs, distribution of a variable.

Once the data collection stage of a study is finished, the Exploratory Analysis (EA) is the first phase of the statistical analysis prior to the descriptive and inferential analysis. EA allows to evaluate the quality of the data collected and written, see if it is possible to correct erroneous data or take into account for later analysis, safeguarding an adequate report of results and conclusions. On the other hand, in the case of quantitative variables, EA allows to evaluate the distribution of these variables. In the article Role and definition of variables in an investigation: the prominence they deserve (1), the reader can review what a quantitative variable is.

There are different graph tools to study the distribution of

Correspondence:
Mauricio Fuentes A., Biostatistics Program, School of Public Health, Universidad de Chile. Independencia 1027, Independencia, Santiago, Chile.
+562 29786554.
mauriciofuentes@med.uchile.cl

quantitative variables, which are presented in the following sections of this article. To illustrate these tools, we used a database, presented by Hosmer and Lemeshow (2), of 189 newborns and their mothers treated at the Baystate Medical Center in Springfield, Massachusetts, USA (available at ftp: // ftp.wiley.com/public/sci_tech_med/logistic). These data were collected within a study whose objective was to investigate whether some characteristics or behaviors during pregnancy (feeding, smoking, prenatal medical care, among others) influenced the weight of the newborn. Among the variables recorded, which will be used in this article, are the following:

- Mother's weight at the date of her last menstrual period, in kilograms (kg).
- Mother's race (white, black, other).
- Mother's smoking habit (smokes, does not smoke).
- Birth weight, in grams (g).

The variable of greatest interest in the study was birth weight, so this article will focus mainly on the behavior of this variable, trying to answer through graph tools the questions; how were the weights at birth distributed among all the new born? ? And; was the distribution different, for example, if the mothers smoked or did not smoke?

HISTOGRAMS

The numbers in a database are the values that each variable takes for each individual in the sample. In the case of a quantitative variable, most values can take different values (3). However, the values are usually grouped in intervals, determining the distribution of the variable, which corresponds to the pattern of occurrence observed in a set of values of a variable (4). Of the various ways to visualize these patterns, the graph used par excellence is the histogram, a graph generally with vertical bars where the height of each bar represents the absolute frequency (number of observations) or the percentage of occurrence of the respective value or range of the variable (4,5). Figure 1 shows the histograms of the birth weight variable and its interpretation. In general, the intervals of the analyzed variable are located on the horizontal axis (X axis) of the histogram and the absolute frequency or the percentage of occurrence is located on the vertical axis (Y axis). The symmetry, one of the most important elements that is evaluated in a histogram, corresponds to the shape shown by the set of bars, and a symmetric distribution is considered when an approximation of the same shape is observed towards both sides (left and right) of its central value, for example, similar to a bell. On the contrary, if a more elongated distribution of the bars is observed to the right or to the left, it would be described as an asymmetric distribution. Another element that can be observed in the histogram is the modal interval or more frequent interval, which corresponds to the highest bar, which is usually unique. However, sometimes more than one modal interval can be found.

Figure 2 compares the histograms of the birth weight of the children of smoking and non-smoking mothers using the absolute frequency and the percentage of occurrence. The comparison of two histograms represented with absolute frequencies should be done with caution, especially if the number of subjects in both groups is different. In this case, it is more appropriate to compare histograms that represent the percentages of occurrence. An important aspect to highlight is the need for the compared histograms to have the same scales, both in the range of the values used in the vertical axis and in the intervals used in the horizontal axis. In this way, the graphic visualization will deliver a correct impression of the information, which is relevant since a graph should allow for a quick interpretation at a first glance. Figure 3 compares the histograms of the weight at the date of the last menstrual period according to the race of the mothers, using different scales (a) and the same scales on the axes (b).

Figure 1. Histograms of the birth weight variable, whose values are shown on the horizontal axis (X axis), and the absolute frequency (number of cases) on the vertical axis (Y axis) or (b) the percentage of occurrence. In both graphs it is observed that the birth weights have a more or less symmetrical distribution, the modal interval (bar of greater height) is 3000 to 3500 g with most of the newborns weighing between 2000 and 4000 g.

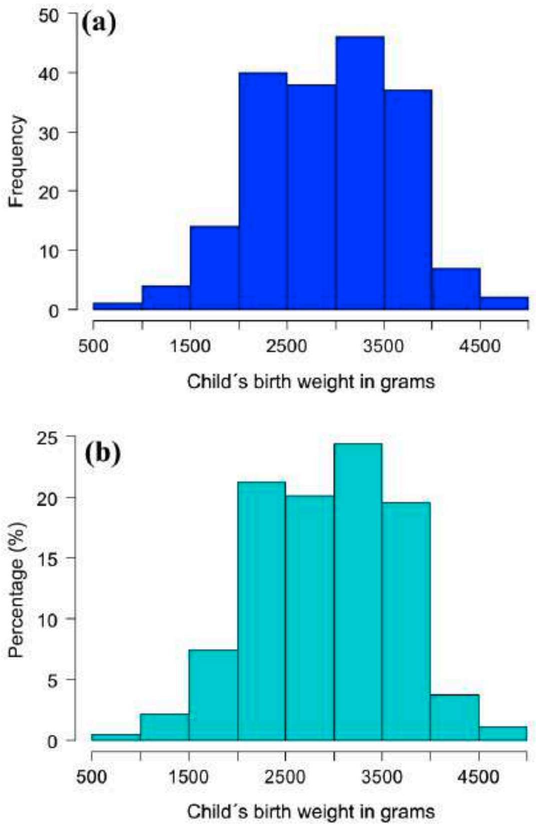


Figure 2. Histograms of the birth weight variable according to the mother's smoking habits, using (a) absolute frequencies and (b) percentage of occurrence. In the histograms of absolute frequencies (a), it is observed that the modal interval of birth weight in the children of smoking mothers was 2000-2500 g, with approximately 20 cases, while in the children of non-smoking mothers the more frequent interval was 3000-3500 g, with about 30 cases. However, it is difficult to know the relative importance of these two values. On the other hand, in the histograms of percentage of occurrence (b) it can be observed that the modal interval in the group of smoking mothers (2000-2500 g), represents approximately 30% of the cases, while the modal interval of the group of non-smoking mothers (3000-3500 g) represents a percentage lower than 30%. In addition, in these graphs the weights are no longer shown as symmetrical as in the complete sample.

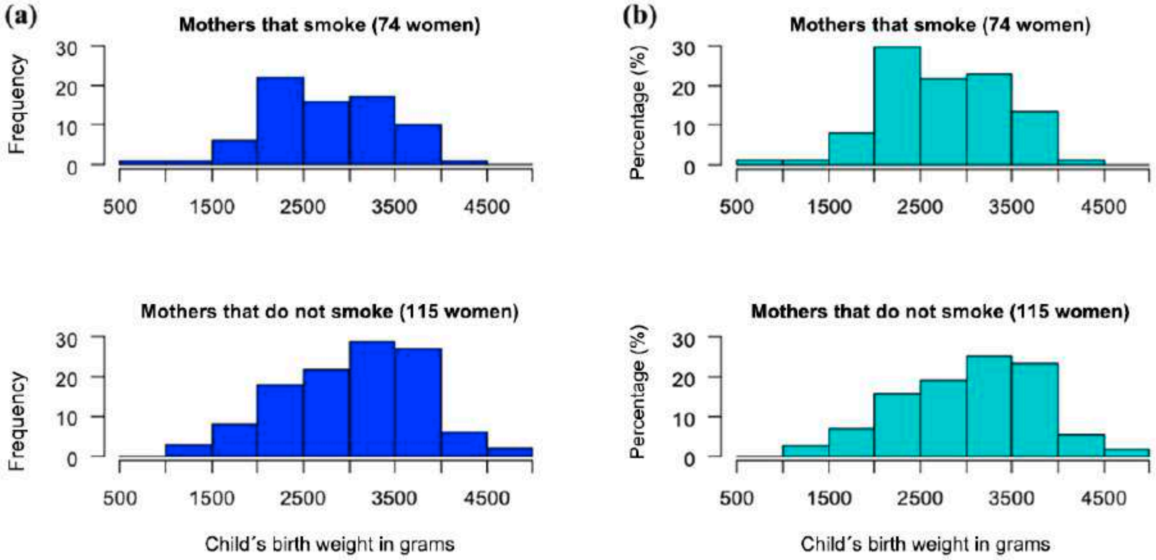
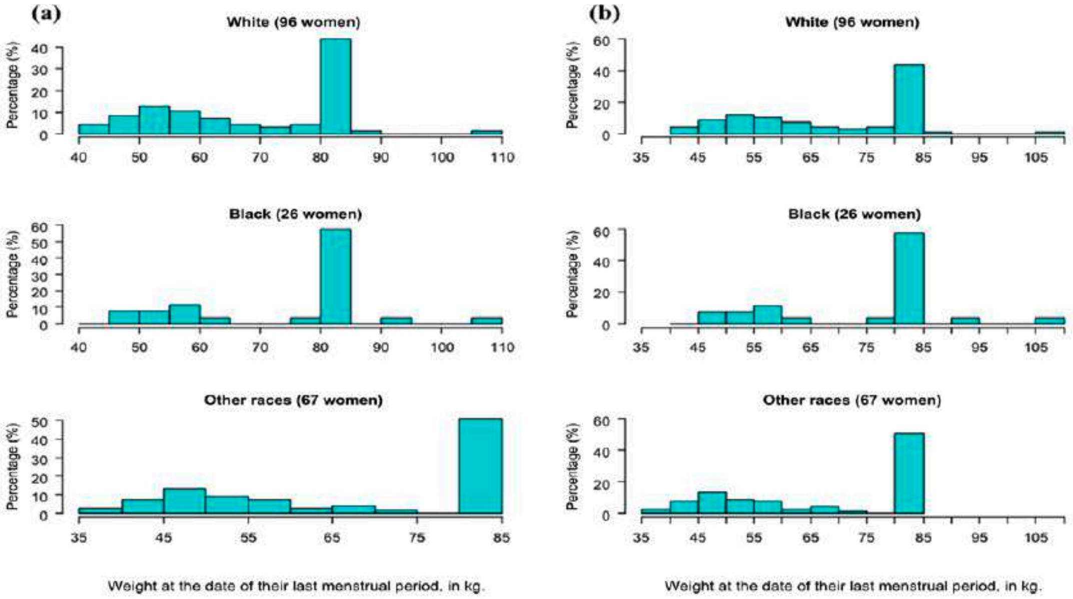


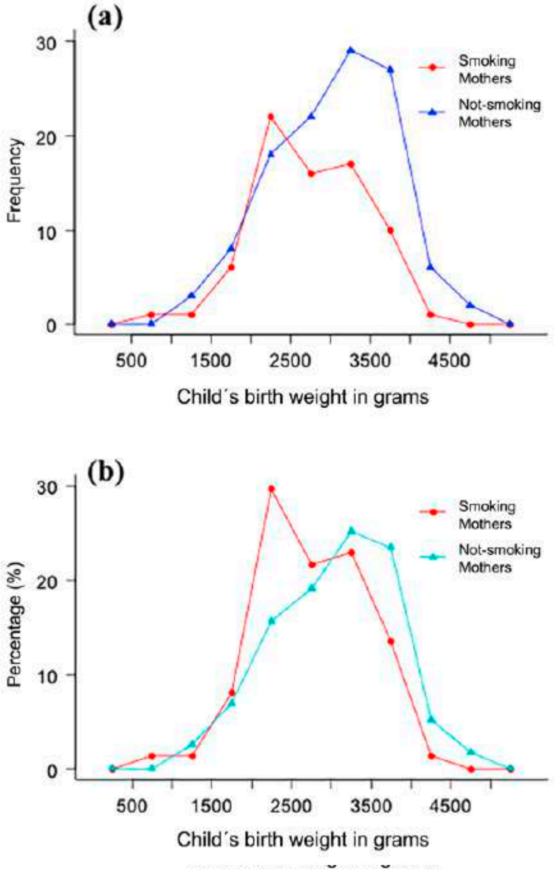
Figure 3. Histograms of the woman's weights at the date of their last menstrual period according to their race using the percentages of occurrence: (a) with different scales on the axes and (b) with the same scale on the axes. In (a), a first impression might indicate that the modal interval is greater in mothers of other races, since in the third histogram the highest bar is more to the right. However, this misinterpretation is product of having used a different scale on the horizontal axis of the last graph. Another misleading impression would be that the modal intervals are similar in the three groups (similar bar heights), however, the vertical axes have different scales. In (b) these aspects were corrected, where it is observed that the modal interval is the same in the three race categories (80-85 kg) with a smaller proportion in white mothers. Other information observed is that no mother in the third group had a weight greater than 85 kg, and that no white or black mother had a weight less than 40 kg.



FREQUENCY POLYGONS

Frequency polygons relate the values of the variable with its respective frequencies, showing the distribution of quantitative data as a series of points connected by means of straight lines, being represented by a curve that turns out to be very useful both to describe the data (5,6) and to compare two or more groups. Similar to that mentioned in histograms, for comparison, it is more appropriate to use occurrence percentages. Figure 4 shows the birth weights using frequency polygons, comparing the group of smoking mothers with that of non-smokers.

Figure 4. Frequency polygons of birth weights for smoking and non-smoking mothers using (a) absolute frequencies and (b) percentages. This type of graph is equivalent to superimposing two histograms, but instead of representing the bars, their heights are joined at the midpoint (class mark of the respective interval) through straight lines. In this way, it is easy to see that both distributions have a similar dispersion, although the children of smoking mothers tended to have lower birth weights. As in Figure 2, the heights of the graphs are different if frequencies or percentages are represented, the second option to compare being better.



BOX-AND-WHISKER PLOT

The box-and-whisker plot is a graph commonly used to compare distributions between groups, even more than histograms and frequency polygons. They are very useful for visually summarizing the shape of a distribution and its degree of symmetry (4). As can be seen in Figure 5, the box shows the positions of the 25th percentiles (lower end), 50th (inner line) and 75th (upper end), which correspond to the quartiles of the distribution or those values that divide the set of data in four equal parts. Particularly, the 50th percentile or second quartile corresponds to the median, a value under which half of the data is found. The lower (25th percentile) and upper (75th percentile) ends of the box indicate that it contains the central half of the values of the observations, which is known as the interquartile range of distribution (4,5). The whiskers extend, in both directions, to the most extreme data which should not be extended more than 1.5 times the interquartile range from the respective edge of the box (7). Any data outside this range, that is, beyond the whiskers, is called "outlier" and is shown in the graph (figure 5). These out-of-range values must be interpreted in the context of each variable, identifying whether it corresponds to a plausible value or registration or coding errors. In the latter case, the researcher may correct, if possible, or take these values into account to make decisions regarding what to do with them in subsequent analyzes.

Some summary measures such as median and percentiles have been mentioned and explained briefly, which will be discussed in depth in an upcoming article in this series. However, even if these concepts are not remembered or fully managed, a great advantage of the box-and-whisker plot for different groups is that they allow to visually identify if there are apparent differences between them.

Another characteristic of the data that is easy to visualize in a box-and-whisker plot is its degree of symmetry. When the median is located near the middle of the box, we can say that the central 50% of the data is approximately symmetric, however if the median is close to the upper edge of the box (75th percentile) or the lower (25th percentile)), the distribution is most certain to be asymmetric (4). A similar interpretation is made in regards to the distance of the whiskers from the box.

In the box-and-whisker plot it is also possible to easily observe the dispersion of the variable. The longer the box is, the wider the range in which the central half of the data is located (interquartile range). Also, the wider the range between the two whiskers, the greater the dispersion or total variability of the data. Evaluating the degree of dispersion of a variable is important in statistical analysis, since it indicates how homogeneous (less dispersion) or heterogeneous (greater dispersion) is the sample in regards to said variable. When a group is more homogeneous, individuals are more similar to each other (at least in the variable of interest), and therefore it is easier to describe and generalize their characteristics.

Figure 5. Comparison of birth weights between smoking and non-smoking mothers using the box-and-whisker plot. The quantitative variable (birth weight) is located on the vertical axis and the groups to be compared (categories of the qualitative variable) on the horizontal axis. The line located inside the boxes indicates the median of each group, these being approximately 2700 g in smoking mothers and 3100 g in non-smoking mothers. The lower (25th percentile) and upper (75th percentile) of the boxes allows estimating that half of the children of smoking mothers were born with weights between 2300 and 3200 g, while half of the children of non-smoking mothers were born with weights between 2500 and 3600 g. The point under the lower whisker of the group of smoking mothers corresponds to an "outlier", and represents a baby born with an unusually low weight (approximately 700 g!). In general, it is observed that the children of smoking mothers had lower birth weights than those of non-smoking mothers. Finally, the weights at birth in the group of smoking mothers were somewhat more asymmetric than the weights of the other group, since the lower whisker is farther from the box than the upper whisker. In addition, it is seen that the extent of both the box and the whiskers in the group of smoking mothers is smaller, which indicates that it has less dispersion than the other group, that is, it is more homogeneous in terms of birth weight.

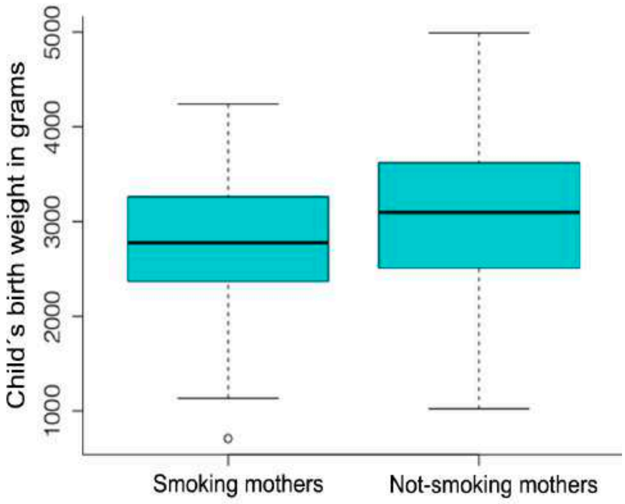
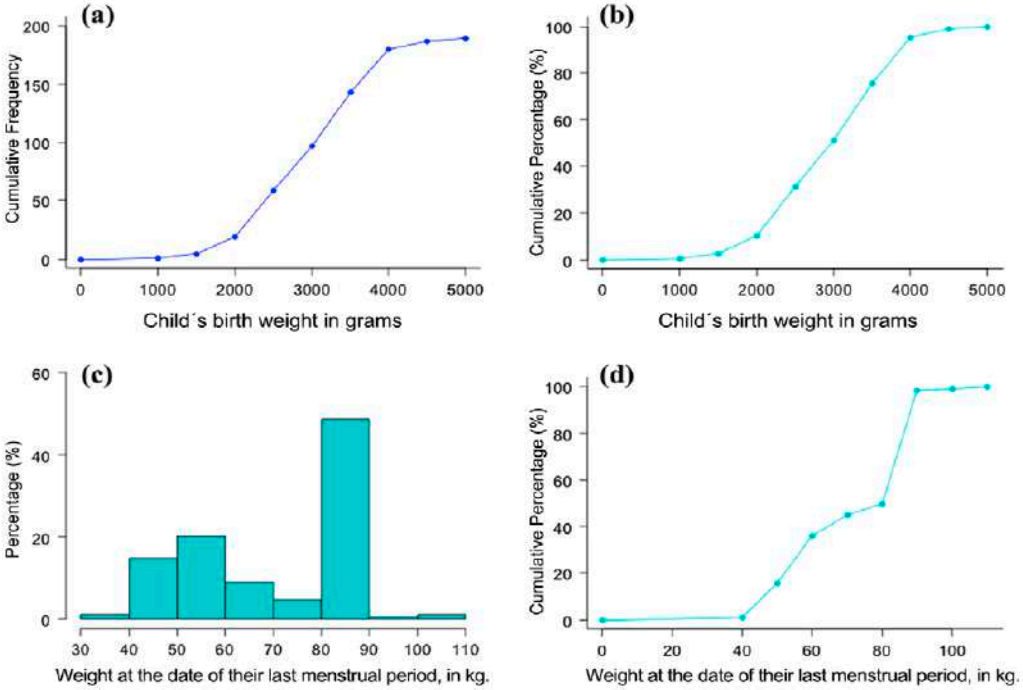


Figure 6. Above: Cumulative frequency polygon or ogive of birth weights using (a) absolute frequencies and (b) percentages. Below: Percentage distribution of mothers' weights at the date of their last menstrual period: (c) histogram, (d) ogive. In (a) each point indicates how many children were born with a weight less than or equal to that value. For example, the sixth point of the curve indicates that (approximately) 100 children had a birth weight less than or equal to 3000 g, which is equivalent to saying that (approximately) 90 children weighed 3000 g or more. In (b) each point indicates the percentage of children born with a weight less than or equal to that value, that is, the sixth point indicates that approximately 50% of children were born with 3000 g or less. Note that including or not the punctual value of 3000 g in one or another portion of the distribution is statistically irrelevant, that is, it does not matter whether half weighed 3000 g or less or to detail that they weighed less than 3000 g. In (c) it is observed that the distribution of the mother's weight at the date of her last menstrual period is far from being symmetrical, and its corresponding ogive in (d) has a more staggered shape, with a pronounced change in the modal value.



ACCUMULATED DISTRIBUTION

Another important aspect in data exploration is knowing the cumulative frequency distribution or cumulative percentage. For a given value of the variable, the cumulative frequency indicates the number of cases with values less than or equal to that value, and the cumulative percentage indicates the percentage of cases with values less than or equal to that number. This can be visualized through a cumulative frequency polygon or ogive as shown in Figure 6. The shape of the ogive depends on the symmetry of the distribution, which is illustrated by comparing ogives (a) and (b) with the (d) one from Figure 6.

CONCLUSION

The most frequently used statistical graphs have been shown and which, according to the authors, are the most useful for exploring quantitative data. The distribution of all the values of a quantitative variable (univariate analysis), as well as the values separated into categories of a qualitative variable (bivariate analysis), were visualized, trying to envision if there is a relationship between the quantitative variable of interest and that qualitative variable. These simple procedures, among other similar ones, allow for the identification in registration and coding errors, increasing the quality and knowledge of the data on which the following analyzes will be based. They also allow to evaluate the distribution of quantitative variables, which inevitably constitutes the basis of the following statistical analyzes, both the descriptive and the use of statistical inference methods (parameter estimation, hypothesis testing, regression analysis, among others), since these consider the distribution of the variable for its application. The omission of this first stage can lead to the report of biased results and erroneous conclusions.

The authors declare no conflicts of interest.

REFERENCES

1. Yohannessen K, Fuentes M. Role and definition of variables in an investigation: the prominence they deserve. Neumol Pediatr [Internet]. 2019; 14 (3): 122–5. Available from: https://www.neumologia-pediatrica.cl/wp-content/uploads/2019/10/1.pdf
2. Hosmer DW, Lemeshow S. Applied Logistic Regression. 2nd ed. New York: John Wiley & Sons, Inc .; 2000
3. Argimon Pallás J, Jiménez Villa J. Métodos de investigación clínica y epidemiológica. (Clinical and epidemiological research methods.) Elsevier España, S.L. 2013
4. Dawson GF. Easy interpretation of biostatistics. The connection between evidence and medical decisions. 1st Edition. Spain: Elsevier España S. L .; 2009
5. Hernández Sampieri R. Metodología de la Investigación (Research methodology.) McGraw-Hill / Interamerican Editors. 2014.
6. Bennett JO, Briggs WL, Triola MF. Statistical Reasoning 1st ed. Mexico: Pearson Educación de México, S.A .; 2011
7. Taucher E. Bioestadistica. (Biostatistics.) Third ed. Santiago, Chile: Ocho libros editores; 2014.