Describing numerical variables: which are the most appropriate parameters to describe the data?*

Rodrigo Pereira Duquia [1]
João Luiz Bastos [3,4]
Renan Rangel Bonamigo [1,6]
David Alejandro González-Chica [2]
Jeovany Martínez-Mesa [5]

DOI: http://dx.doi.org/10.1590/abd1806-4841.20174904

Abstract: The proper description of numerical variables is very important when presenting a set of data. Measures of central tendency and dispersion are used to adequately understand a set of numerical variables. Knowledge of the properties of these measures and their adequate use provide the reader with a better understanding of the results of a study.

Keywords: Analytical epidemiology; Epidemiology; Epidemiology, descriptive

INTRODUCTION

When we plan a study, an essential step is to think about how to measure the variables of interest. Will the measurements be taken on a continuous or categorical scale, for example? The level of detail, as well as the quality of the data collected are key aspects when it comes to measuring variables with adequate levels of validity and reliability. If data quality is poor, the analyses will not provide valid nor reliable results – for a more detailed discussion of validity and reliability in scientific studies, we refer the reader to a previous article of this series.[1]

After collecting data for a study, constructing the corresponding data set, checking and codifying the variables, the following step is to describe how the variables are distributed. The description of the study variables should be made according to the research question, the study design and the type of variable in question (categorical/qualitative or numerical/quantitative, as described in the previous article).[2] In this article, we approach the main parameters used to describe numerical variables, and also review some basic principles of data distribution that will help the reader decide which parameters are more appropriate for their assessment.

When working with numerical variables (whether discrete or continuous), such as total daily sun exposure time, blood markers levels, or size of skin lesions, one must go beyond the presentation of observed data for each individual or unit of analysis; it is critical to make use of resources that allow for a broader assessment of the data set, beyond individual observations.[3-5] Some summary measures indicate the most common values in a data set – these are commonly known as measures of central tendency (MCT) and usually reflect the most frequent values in a group of individual observations.[4,5] On the other hand, it is also necessary to know how dispersed these data are in relation to the most frequent ones; in this case, we use measures of dispersion or data variability (MD).[4,5]

For example, a dermatologist who would like to evaluate laboratory data of a sample of 20 patients on methotrexate for over 6 months, decides to estimate their levels of alanine aminotransferase (ALT) and of C-reactive protein (CRP). The data for each patient are described in table 1. In order to provide a MCT for the 2 numerical variables (ALT and CRP), i.e., provide a summary measure that reflects the values of the 20 observations, the researcher could choose the mean, median, or mode. Below, we will see these different MCTs, their interpretations, and implications based on the analysis of the set of 20 cases.

MEAN

Mean is considered a parametric measure, for it is based on the specific values of each individual and, therefore, is one of the most commonly used MCT.[4-6] Mean is calculated by the sum of the individual values, divided by the number of observations (n).[3-6] Using the data from table 1 to calculate the mean for ALT and CRP in the aforementioned group of patients, the corresponding values are 30.9 and 6.5, respectively.

Received on 05.07.2015.
Approved by the Advisory Board and accepted for publication on 07.07.2017.

* Study conducted at Universidade Federal de Ciências da Saúde de Porto Alegre (UFCSPA) – Porto Alegre - RS, Brazil.

Financial support: none.
Conflict of interest: none.

[1] Department of Dermatology at Universidade Federal de Ciências da Saúde de Porto Alegre (UFCSPA) – Porto Alegre - RS, Brazil.
[2] Discipline of General Practice, School of Medicine, The University of Adelaide – Adelaide, Australia.
[3] Department of Public Health at Universidade Federal de Santa Catarina (UFSC) – Florianópolis - SC, Brazil.
[4] Post-graduate Program in Collective Health of the Universidade Federal de Santa Catarina (UFSC) – Florianópolis - SC, Brazil.
[5] School of Medicine of the Faculdade Meridional (IMED) – Passo Fundo - RS, Brazil.
[6] Post-graduate Program in Pathology of the Universidade Federal de Ciências da Saúde de Porto Alegre (UFCSPA) – Porto Alegre - RS, Brazil.

MEDIAN

Median is considered a non-parametric measure, since individual data are organized from the lowest to the highest value (*ranking*) and, regardless of each individual value, the one located in the position that divides the data set in two equal parts is selected.[4,5] Therefore, there will always be 50% of the observations below the median and 50% above it. The median is also known as 50th percentile or simply P50.

To identify the position that will indicate the median value, we must use the formula (n+1)/2, where *n* represents the number of individuals included in the analysis.[3,6] When the number of individuals is odd, the result of the equation will indicate the exact position of the median value. However, when the number of observations is even, as with the data in table 2 (n = 20), the results of the equation will be a fraction (10.5 in this case). Given that this is a position in a *ranking* of ordered values and since there is no observation in the position 10.5, the median is estimated based on the mean position of the values located in the positions 10 and 11 – for ALT (30+30)/2=30.0 and for CRP (4+4)/2=4.

Table 1: Alanine aminotransferase (ALT) and C-reactive protein (CRP) values in methotrexate users for longer than 6 months

User number	ALT values	CRP values
1	30	5
2	25	4
3	45	38
4	35	2
5	40	10
6	40	6
7	45	16
8	30	4
9	25	4
10	25	1
11	30	3
12	30	5
13	35	7
14	18	1
15	20	3
16	25	3
17	30	4
18	20	3
19	35	5
20	35	6

Fictional data

MODE

Mode is another non-parametric MCT that is less used in clinical practice. It only provides the information of which value more frequently occurs in the data.[3-6] For the ALT data, the mode is 30 (it appears 5 times), but for CRP we have a bimodal distribution, since the values 3 and 4 appear the same number of times (4 times each). There are cases in which mode does not exist, when all possible values have the same frequency, but there can be distributions with 3 or more modes. This is one of the reasons why mode is rarely used.

As observed, by using the different MCTs, we find different values for mean, median and mode for ALT (30.9; 30.0; 30.0, respectively) and for CRP (6.5; 4.0; 3-4, respectively). The decision about which MCT to use will depend on the assessment of symmetry (normality) of the underlying numerical variable. A variable is considered asymmetrical (or non-normal) when the data are dispersed in relation to the MCT or, in other words, when there are too many extreme values in the data.[3-6]

STANDARD DEVIATION

The standard deviation (SD) is a MD, and is one of the parameters that can give you an idea of symmetry in the distribution of variables.[3-6] It estimates the distance of each individual value in relation to the sample mean. The higher the SD, the larger the dispersion of data and, therefore, the higher the likelihood of an asymmetrical distribution. A practical procedure to evaluate the magnitude of SD is to compare it to its mean, using the coefficient of variation (or of dispersion), by means of the equation (SD/mean)*100. The coefficient of variation indicates how high the SD is in relation to the mean. Results above 50% indicate variable asymmetry, and lower percentages (for example, up to 25%) suggest that the variable is symmetrical. In the example from table 1, the SD values for ALT and CRP were 7.9 and 8.1, which, as absolute values, are similar. However, the coefficients of variation were 25% and 125%, which suggest symmetry for ALT and asymmetry for CRP.

SYMMETRY

Symmetry can also be assessed by comparing the values of the mean and the median, which are similar in the case of symmetrical variables and very different from each other in the case of asymmetrical variables. In the case of ALT, both parameters are close to each other (<10% of difference between mean and median), while for CRP the mean was higher than the median (63% higher), suggesting, once again, asymmetry for this latter variable. In addition, other parameters can be used in the evaluation of symmetry, among them:

- *Frequency histogram, a graphical way to assess asymmetry*.[3-6] In the case of figure 1, ALT shows a similar pattern to what is consid-

Table 2: Median values of alanine aminotransferase (ALT) and C-reactive protein (CRP) values in methotrexate users for longer than 6 months

	Ranking of the users of both medications																	
Test		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Individual test values	**Alanine aminotransferase (ALT)**	18	20	20	25	25	25	25	30	30	30	30	30	35	35	35	35	40
	C-reactive protein (CRP)	1	1	2	3	3	3	3	4	4	4	4	5	5	5	6	6	7

Fictional data

ered a symmetrical variable, but for CRP the figure suggests asymmetry (many individuals concentrated in a specific range of values and a small amount of them with extreme values);

- *Skewness or coefficient of asymmetry*, which shows whether individual values are well balanced on both sides of the mean (skewness = 0), or if there are more values to the right (>0) or to the left (<0) of the mean.[3-6] The more distant from zero the skewness is (for example, >+3 or <-3), the higher the likelihood of asymmetry (in the example, ALT = 0.2; CRP = 3.2);

- *Kurtosis*, which takes into account both the dispersion and the shape of the variable distribution, assigning a value of 3 to normal distributions; the higher the value (for example, >10), the higher the likelihood of asymmetry (in the example, ALT = 2.2; CRP = 12.6); and [3-6]

- *Statistical tests for asymmetry (such as Shapiro-Wilk or Shapiro-Franca tests)*, which should be used with caution in large samples; small deviations from normality may be reflected in a statistically significant result (p<0.05), even when the other parameters may suggest symmetry.[3-6]

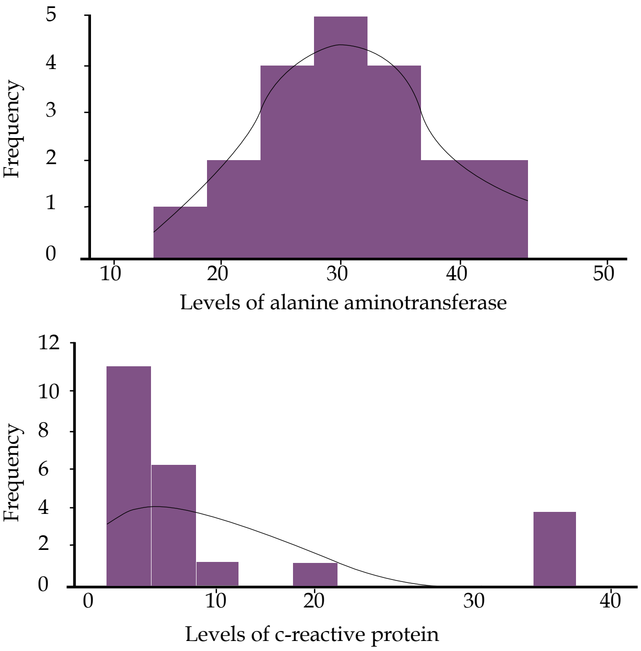

Figure 1: Frequency histograms of the levels of alanine aminotransferase (ALT) and C-reactive protein (CRP) in a sample with 20 patients

When the variable is symmetrical, such as the case of ALT, although the mean and the median are very similar and can be indistinctly applied as an MCT, it is more appropriate to use the mean because it is the most used parametric data. Because the MCT should be accompanied of a MD, we can also use the SD to describe the ALT values, since symmetry of this variable suggests that SD is an adequate measure to assess the dispersion of these data.

For CRP data, where more than one parameter points towards asymmetry, the use of the mean and SD would be incorrect. According to the theory of normal distribution, 50% of the individuals assessed (n=10) should have a value above the mean for CRP[3,5,6]. However, according to the data presented in table 1, only 4 individuals are above this value. In addition, considering the same theory of normal distribution, 95% of the sample should be included in the values corresponding to the mean ± 2 SD (between -10.2 and +22.2 for CRP). Due to the fact that the SD is very high (8.1) in relation to the mean CRP, the lower limit (-10.2) includes negative values for this variable, which is incompatible with the individual values shown in table 1. Other parameters estimated based on the mean and SD, such as confidence intervals or even parametric statistical tests, are not recommended in the case of asymmetric variables. Therefore, the correct MCT for CRP will be the median, which will not be influenced by extreme observations. Like MD, it will be more appropriate to use the total range (difference between the minimum and maximum values) or the interquartile range.

CONCLUSION

MTCs and MDs are widely used in publications in the medical field and readers often do not fully understand their properties. Understanding these measures is essential for readers of scientific articles. In addition to the uses described above, these measures allow for the calculation of sample sizes, as well as assessing whether statistical tests were correctly used: many statistical tests assume a symmetrical distribution of values, such that their use with asymmetric data produces invalid results[3-6]❑

REFERENCES

1. Bastos JL, Duquia RP, González-Chica DA, Mesa JM, Bonamigo RR. Field work I: selecting the instrument for data collection. An Bras Dermatol. 2014;89:918-23.
2. Duquia RP, Bastos JL, Bonamigo RR, González-Chica DA, Martínez-Mesa J. Presenting data in tables and charts An Bras Dermatol. 2014;89:280-5.
3. Altman DG. Practical statistics for medical research. London: Chapman & Hall; 1997.
4. Hayslett HT. Statistics. Oxford: Elsevier; 2014.
5. Kirkwood BR, Sterne JAC. Essential medical statistics. Malden: Blackwell Science; 2003.
6. Maroco J, Bispo R. Estatística aplicada às ciências sociais e humanas. Lisboa: Climepsi Editores; 2005.

Mailing address:
Rodrigo Pereira Duquia
Rua Independência, 172 - sala 902
Centro
90035-070 Porto Alegre, RS
Brazil
E-mail: rodrigoduquia@gmail.com

How to cite this article: Duquia RP, González-Chica DA, Bastos JL, Martínez-Mesa J, Bonamigo RR. Analyzing numerical variables: which are the most appropriate parameters to describe the data? An Bras Dermatol. 2017;92(6): 841-3.