Journal of Community Hospital Internal Medicine Perspectives

co-action PUBLISHING

BIOSTATISTICAL CONCEPTS AND TOPICS IN RESEARCH

Central tendency and variability in biological systems

Lucien J. Cardinal, MD*

Internal Medicine Residency Program, Department of Medicine, Stony Brook Medicine – Mather Hospital, Port Jefferson, NY, USA

In this article, the author reviews the manner in which researchers characterize data. Normality, standard deviation, mean, and other concepts related to parametric statistics are discussed in common language, with a minimum of jargon and with clinical examples.

Keywords: *mean*; *median*; *mode*; *average*; *normal distribution*; *standard deviation*; *p-value*; *statistics*

*Correspondence to: Lucien J. Cardinal, Internal Medicine Residency Program, Department of Medicine, Stony Brook Medicine – Mather Hospital, 75 North Country Road, Port Jefferson, NY 11777, USA, Email: LCardinal@matherhospital.org

Received: 20 March 2015; Revised: 13 April 2015; Accepted: 20 April 2015; Published: 15 June 2015

This article is one of two parts covering central tendency and variability. The second part will be in a subsequent issue of this publication and will cover normal distribution and *p*-value. Usually, when a set of values of a particular biologic characteristic are examined, the group of values can be characterized precisely and reproducibly in terms of two defining features: the center and a measure of how closely values cluster around the center. The foundation of many statistical techniques seen in the medical literature is based on the aforementioned simple concept applied in a sophisticated manner (e.g., any analysis referencing *p* value, standard deviation (SD), analysis of variance, or *t*-test) (1, 2, 3). The terms average, arithmetic mean, median, and mode are all related to central tendency and each defines a concept of center (4). Similarly, the terms range, SD, standard error, dispersion, error, and noise each defines or characterizes how a group of values is scattered around the center.

We all recognize that there is variability in biological characteristics, e.g., height of individuals, head circumference, and left ventricular stroke volume (5, 6). The human body is a complex, organized biological system. A review of medical research publications reveals countless examples of characteristics that are measured and given a value. Each characteristic can be described, precisely and reproducibly by two derived values, one representing a stable center point and the other the degree of variation around the center.

There are two points to keep in mind when reading this article. First, the general concepts described apply primarily to normally distributed values. Fortunately, many biological variables are normally distributed. Additionally, statistics based on the assumption of normality are encountered frequently in the medical literature. Secondly, values must be randomly selected if one is to apply the concepts described. We will examine the meaning of 'normal distribution' in the second part of this article and the concept of random selection of samples in a later article.

Throughout this article, we will use the example of the red blood corpuscle (RBC) volume to illustrate the presented statistical concepts. The practicing clinician is familiar with the complete blood count (CBC), the average (mean) RBC volume (MCV), and the red cell volume distribution width (RDW). RBCs in humans commonly have an average volume of 80–100 femtoliters (fL), which is considered the normal range. Because size and shape may vary (e.g., ovalocyte, spherocyte, and typical biconcave disc) and the RBC occupies a three-dimensional space, characterization in terms of volume may be preferable to characterization by diameter. Values less than 80 fL are considered microcytic and values greater than 100 fL are considered macrocytic.

Central tendency

When a group of biological values is examined, it is common to observe the following: there is variability, the

Citation: Journal of Community Hospital Internal Medicine Perspectives 2015, **5**: 27930 - http://dx.doi.org/10.3402/jchimp.v5.27930

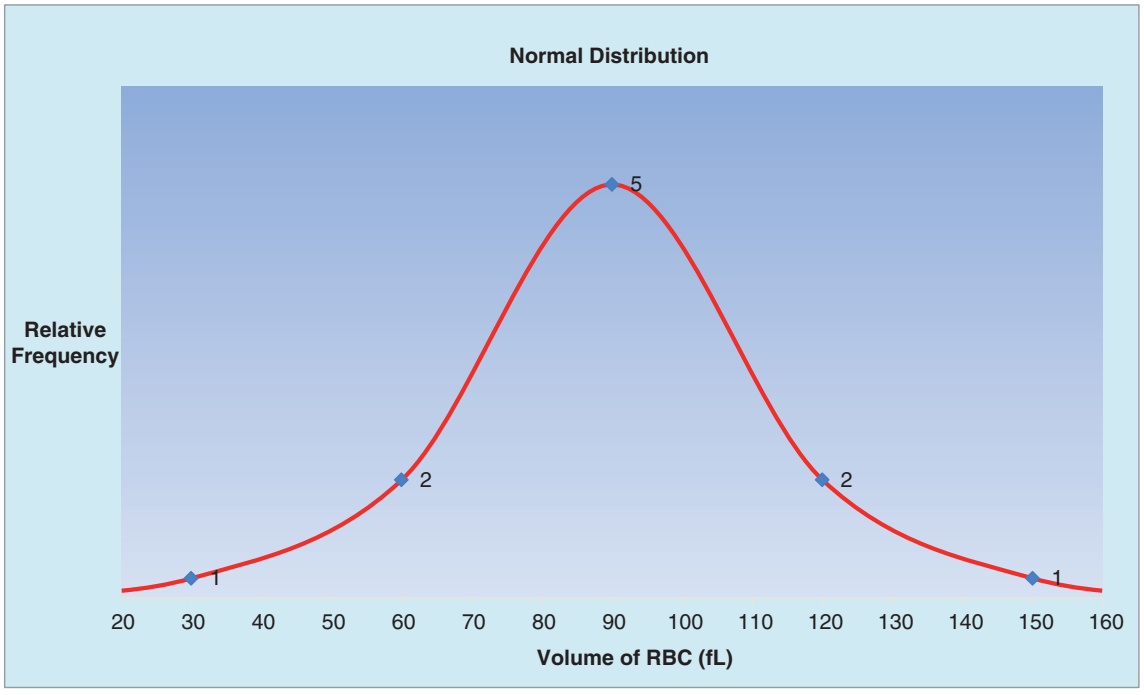

Fig. 1. A common pattern of symmetric distribution. The numbers adjacent to the blue diamonds indicate the relative frequency of RBCs of any given volume.

midway point between the high and low values corresponds to the calculated average value, and the calculated average value corresponds to the most frequently observed value. These observations are related to the often-cited mean, median, and mode. The calculated average (sum of values divided by the number of values) is referred to as the mean. The middle value of a series arranged from least to greatest is referred to as the median (e.g., 2 is the median of the series 1, 2, 3), and the most frequently observed value is the mode (e.g., 2 is the mode of the series 1, 2, 2, 3). The mean, median, and mode are commonly observed to be approximately equal (Fig. 1). The fact that the median value is the middle value in a series corresponds, additionally, to the fact that half of all values will lie below the center and half will lie above the center. The MCV is a measure of the average RBC volume.

As an example of the above topic, if you consider the cell volumes of a large number of RBCs from a single specimen, with a calculated average volume of 90 fL (MCV), you will expect to find variation between the volumes of RBCs. About half of the RBCs will have a volume below the MCV and vice versa (the mean is equal to the median). And the most frequently observed RBC volume will be at or close to the MCV (the mean is equal to the mode).

The central value tends to be predictably stable and can be used as a reliable identifying characteristic. For example, if one looked at two blood specimens and the MCVs were markedly different, it would be correct to surmise that the specimens were likely from different patients or from the same patient at two different points in time (e.g., before and after transfusion) (7).

Variability

The more variable a specimen is, the further values tend to be spread away from the center and, hence, are more dispersed. Variability may be referred to as dispersion. Because the center may be considered as the true target value, variability may be referred to as error, buzz, or noise. When a sample of values is thought to represent a normal distribution, the preferred measure of variability is the 'standard deviation'. As the SD increases so does the variability.

It is useful to know the variability of measured values. Variability, in this context, is a measure of how close the values lie to the center. Measuring the distance from the least to the greatest in a group of values is a simple and useful measure of variability. This distance or spread of values is referred to as the 'range'. Consider this example of two blood specimens, each having an MCV of 90 fL. Specimen 1 has RBCs with volumes ranging between 30 and 150 fL. Specimen 2 has RBCs with volumes ranging between 70 and 110 fL. The difference between the cell of greatest and least volume is 120 fL (150 − 30) in specimen 1 and 40 fL (110 − 70) in specimen 2. The volumes of RBCs in specimen 2 tend to be closer to the MCV; hence, specimen 2 demonstrates less variability compared to specimen 1 (Fig. 2). Comparing Figs. 1 and 2,

Citation: Journal of Community Hospital Internal Medicine Perspectives 2015, **5**: 27930 - http://dx.doi.org/10.3402/jchimp.v5.27930

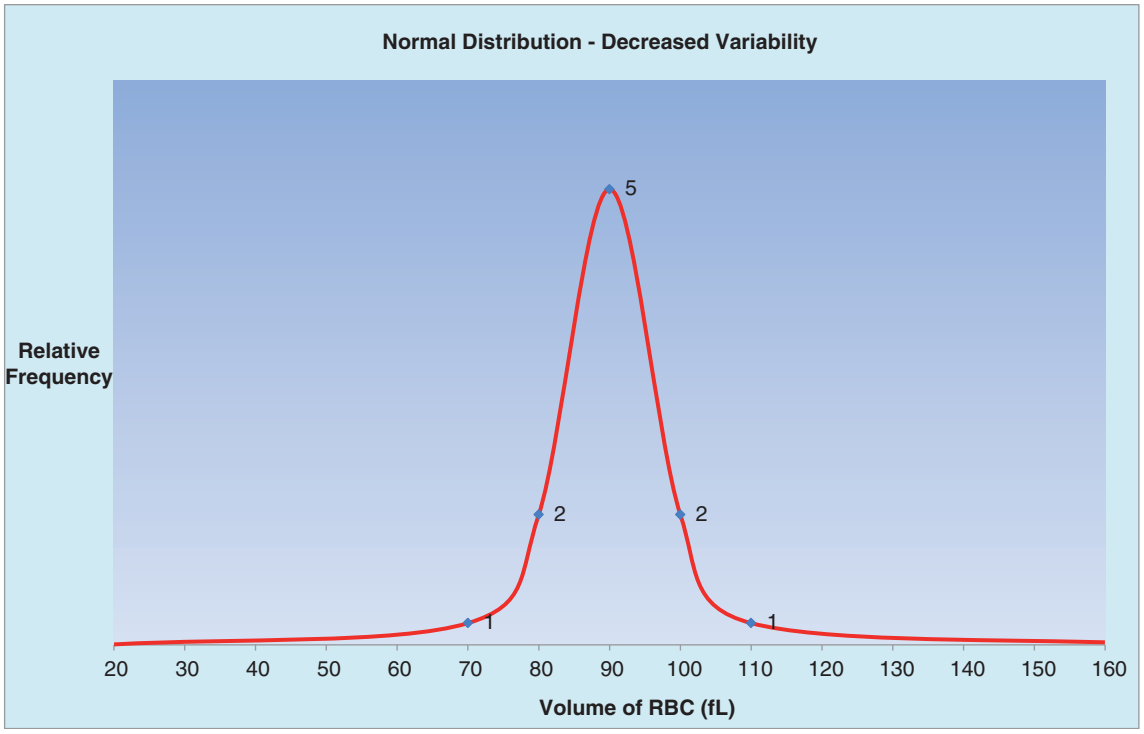

Fig. 2. Two groups may have the same center and yet be different. When examining such figures as Fig. 2, note that the wider the curve (the bell), the more variable the population. The RBC specimen represented in this figure has the same average volume (90 fL) as Fig. 1; however, it demonstrates less variability. The fact that the variability is less in Fig. 2 means that the SD is also less, since the SD is a measure of variability. Note that the range in Fig. 1 is about 120 fL (30–150) and in Fig. 2 is about 40 fL (70–110). Variation is a stable and useful identifier. 'Normal distribution' will be defined in the second part of this article.

it is easy to see that the graphical representation appears wider as the variability increases. Variability, like the central value, is a precise and reproducible identifying characteristic. Similar to the example provided in Central Tendency, when two blood specimens have markedly different variability in RBC volume, it suggests that the specimens come from different patients, different times, or different circumstances (8, 9).

The CBC characterizes each specimen in terms of the variability of RBC volume. The RDW is the measure of variability reported on the CBC. It is reported as a percent

$$\text{RDW} = \text{SD of RBC volume} \div \text{MCV} \times 100$$

In the example described above, the RDW of specimen 2 would be less than the RDW of specimen 1. The SD is a measure of variability and the RDW incorporates the SD into its calculation. The RDW may be impacted by various disease states and therefore may be useful in the diagnostic evaluation (10, 11).

Main points

Any set of measured values can be characterized in terms of its center and range of values, the range being a measure of variability.

The range is characterized by the span between the least value and the greatest value.

Variability is conveniently characterized by a measure of SD when a set of values is normally distributed.

Acknowledgements

The author gratefully acknowledges the contributions of the following individuals whose thoughtful reviews and constructive criticisms contributed to the completion of this document: Razvan Hurezeanu, MD; Alan T. Kaell, MD; Srinivas Madhavan, MD; and Ayesha Qadir, MD.

Conflict of interest and funding

The author has not received any funding or benefits from industry or elsewhere to conduct this study.

Bibliography

1. Indrayan A. Medical biostatistics. 3rd ed. Boca Raton, FL: CRC Press; 2012.
2. Glantz SA. Primer of biostatistics, 7th ed. New York: McGraw-Hill; 2011.
3. Hebel JR, McCarter RJ. A study guide to epidemiology and biostatistics, 7th ed. Sudbury, MA: Jones & Bartlett Learning; 2012, p. 219.
4. Anderson S. Biostatistics: A computing approach. Boca Raton, FL: CRC Press; 2012, 306 p.

5. Cheng CK, Chan J, Cembrowski GS, van Assendelft OW. Complete blood count reference interval diagrams derived from NHANES III: Stratification by age, sex, and race. Lab Hematol 2004; 10(1): 42–53.
6. Fraser CG, Wilkinson SP, Neville RG, Knox JD, King JF, MacWalter RS. Biologic variation of common hematologic laboratory quantities in the elderly. Am J Clin Pathol 1989; 92(4): 465–70.
7. Bessman JD, Feinstein DI. Quantitative anisocytosis as a discriminant between iron deficiency and thalassemia minor. Blood 1979; 53(2): 288–93.
8. Fraser CG. Analytical goals for haematology tests. Eur J Haematol Suppl 1990; 53: 2–5.
9. Steinberg MH, Dreiling BJ. Microcytosis. Its significance and evaluation. JAMA 1983; 249(1): 85–7.
10. Roberts GT, El Badawi SB. Red blood cell distribution width index in some hematologic diseases. Am J Clin Pathol 1985; 83(2): 222–6.
11. Bergin JJ. Evaluation of anemia. Getting the most out of the MCV, RDW, and other tests. Postgrad Med 1985; 77(8): 253–4, 6, 61–9.