2014

Iuliu Haţieganu University
of Medicine and Pharmacy
Cluj-Napoca

Sorana D. BOLBOACĂ

Descriptive statistics parameters:

Measures of centrality

Contents

Definitions

Descriptive statistics = methods or indicators that allow to represent data in a readable form. Some methods allows to prepare data for graphical representation (such as histogram, bar chart, pie chart, box plot, etc.) while others allow to obtain parameters that summarize the investigated data (such as mean, standard deviation, correlation coefficient, etc.)

Estimation [1] = the procedure used to determine the value of a particular parameter associated to a population. Two main types of estimators are used in medical statistics: point estimation and interval estimation.

Estimator = statistical function of a sample used to estimate an unknown parameter of the population. The value obtain is an estimate of the population parameter.

Measures of centrality = simple values that give us information about the distribution of data such as arithmetic mean, median, mode.

Classification of descriptive statistics parameters

- Measures of central tendency
- Measure of dispersion
- Measures of localization
- Measures of shape

More about central tendency estimators

Arithmetic mean [2] = a measure of central tendency which allows to characterize the center of the frequency distribution of a quantitative variable that follow a normal distribution. It is calculated by summing all numbers in the sample and then dividing by the number of observations. If the distribution of data is symmetric and unimodal the arithmetic mean is equal to both median (M_d) and modal (M_o) values. If the distribution is unimodal $m \geq M_d \geq M_o$ for data skewed to the right and $m \leq M_d \leq M_o$ for data skewed to the left.

Median (M_d) = a measure of central tendency used when data are not normal distributed. Its value is not influence by outlier since just a maximum of two values are implied in the calculation; is easy to determine because only one classification is needed; is easy to understand. It is the estimator of central value of a distribution when it is asymmetric or has outliers. The median split the dataset in the way in which 50% of observations are on each side of its value.

The steps needed to be follow for calculation the median are:

- Arrange the n observations in increasing or decreasing order.
- For n = odd: $M_d = X_{(n+1)/2}$ where M_d = median, X_i = the i^{th} observation. The median equals the value of the middle observation.
- Form n = even: $M_d = X_{n/2} + X_{(n/2+1)}$. The median equals the arithmetic mean of the values of these two observations.

Mode (M_o) = a measure of central tendency defined as the value with the highest frequency (most represented value of a set). The mode has a value that is little influenced by outliers but is a rarely used measure. Its value is strongly influenced by the fluctuations of a sampling and can strongly vary from one sample to another. A distribution can have a unique mode (called the unimodal distribution) or many modes (called the bimodal, trimodal, multimodal distribution).

For a discrete variable, the mode has the advantage of being easy to determine and interpreted and it is used especially when the distribution is not symmetric. For continuous variable, the mode is a good indicator of the center of the data only if there is one dominant value in the data set. If there are many dominant values, the distribution is called plurimodal, cases in which the modes are not the measure of central tendency. A bimodal distribution for example indicates that the considered distribution is in reality heterogenous and is composed of two sub-populations with different central values.

Relationship between central tendency parameters and distribution

The assessment of mean, median and mode could lead to identify the type of asymmetry in the data (see Figures 1-2).

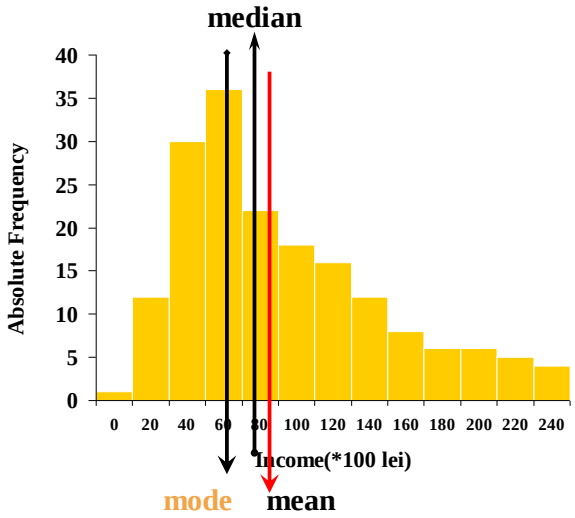

Figure 1. *Left (positive) asymmetry: Mode < Median < Mean*

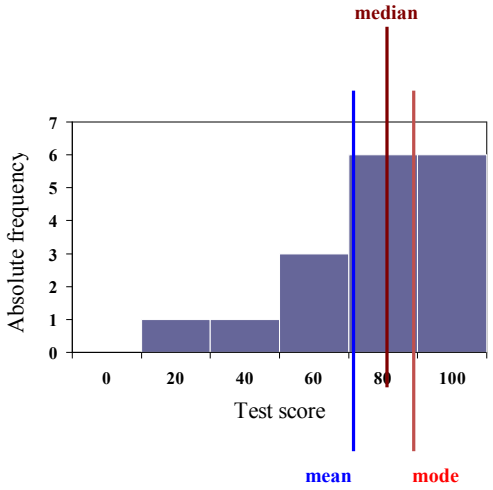

Figure 2. *Right (negative) asymmetry: Mode > Median > Mean*

Evaluating descriptive statistics statements

The evaluation of any statistical parameters and statements must take into consideration both the relevance and the validity of data and analysis on which the statement is based on.

The easiest way to evaluate a statistical statement is to follows the Huff's criteria (5 questions) [3]:

1. *Who says so?*
2. *How does he/she know?* The estimator reflects the facts?
3. *What's missing?* All information needed to proper interpretation was provided?
4. *Did someone change the subject?* The estimator offers the right answer to the wrong problem?
5. *Does it make sense?*

References

[1] Fisher RA. On the mathematical foundations of theoretical statistics. Philos. Trans. Roy. Soc. Lond. Ser A 1922;222:309-368.
[2] Simpson T. Aletter to the RightHonorable George Earl of Macclesfield, President of theRoyalSociety,on the advantage of taking the mean of a number of observations inpractical astronomy. Philos.Trans.Roy.Soc. Lond. 1755;49:82-93.
[3] Huff, D., 1954, How to Lie with Statistics, W. W. Norton & Company, New York, NY.