Received: 14 Nov. 2012 **Accepted: 27 Dec. 2012**

Review Article

Statistics for dental researchers: descriptive statistics

Arash Shahravan DDS, MS[1], Amir Reza Ghassemi DDS[2], Mohammad Reza Baneshi PhD[3]

Abstract

Descriptive statistics is the process of summarizing gathered raw data from a research and creating useful statistics, which help the better understanding of data. According to the types of variables, which consist of qualitative and quantitative variables, some descriptive statistics have been introduced. Frequency percentage is used in qualitative data, and mean, median, mode, standard deviation, standard error, variance, and range are some of the statistics which are used in quantitative data. In health sciences, the majority of continuous variables follow a normal distribution. skewness and kurtosis are two statistics which help to compare a given distribution with the normal distribution.

J Oral Health Oral Epidemiol 2012; 1(2): 53-59

Descriptive statistics is a procedure through which raw data are summarized and simple, yet useful, statistics are generated for the audience.[1]

Suppose a dentist poses the question "What is the average length of maxillary lateral incisors for endodontic treatment?", to this end, the dentist extracts the working lengths of these teeth from 10 patient files and writes them down on paper as follows: 19, 18.5, 21, 21, 18, 19.5, 20.5, 20, 18, 19.

A brief look at the data, even without summarizing, shows that it is possible to gain some insight into the condition of the data, albeit with some difficulty. For example it is possible to say what the maximum and minimum lengths are. However, when there is an increase in the number of items in the data it becomes necessary to carry out some simple calculations to make data more comprehensible. Researchers and statisticians do this by implementing the principles of descriptive statistics. In addition, in big sample sizes it is difficult and time-consuming to report the characteristics of the samples individually in articles, and beyond the scope of time and patience of the audience in lectures. Therefore, if it is possible to summarize data by the use of applied statistics and present them in an article or a lecture, comprehensiveness of statistics makes it possible to compare different groups of data with each other. Descriptive statistics makes it possible to use accurate and scientific statistics in order to summarize data in the best manner possible for the audience. We hope you have now realized the importance of descriptive analysis.

The first step in reporting descriptive data is to determine whether the variable under consideration is qualitative or quantitative, because there are different principles for reporting data of different types of variables. It should be pointed out that a variable is a characteristic of an individual or a phenomenon, which can be measured and can assume different values in a research.[2] For example, in the example given above, the length of the maxillary lateral incisor is the variable under question, because it is measureable and assumes different values. However, the tooth type is not a variable

1- Associate Professor, Kerman Oral and Dental Diseases Research Center, Department of Endodontics, School of Dentistry, Kerman University of Medical Sciences, Kerman, Iran
2- Private Practitioner, Kerman Oral and Dental Diseases Research Center, Kerman University of Medical Sciences, Kerman, Iran
3- Assistant Professor, Research Center for Modeling in Health, Department of Epidemiology and Biostatistics, School of Public Health, Kerman University of Medical Sciences, Kerman, Iran
Correspondence to: Mohammad Reza Baneshi PhD
Email: m_baneshi@kmu.ac.ir

because all the teeth involved are maxillary lateral incisors. As a result, this characteristic does not vary and is constant.

Different Types of Variables

Variables can be divided into two general categories: quantitative and qualitative. Quantitative variables are expressed using numbers; for example, height, weight, the number of children in a family, DMF index (an index for the prevalence of caries), and root length are quantitative variables.

Qualitative variables are classified based on titles. For example, gender (male or female), satisfaction with dental services (very satisfied, fairly satisfied, dissatisfied, very dissatisfied), impression quality (good, moderate, poor), and post-operative pain (severe, moderate, mild, without pain) are qualitative variables. Of course, it should be noted that if the patients are asked to assign a numeric value from 0 to 10 to the severity of pain (VAS), the variable is quantitative.

Description of Qualitative Data

Frequency is used to describe qualitative data, which means the number of individuals with a particular characteristic or trait. Of course, in reporting the frequency of the variable in question, the sample size is important. For example, if in a class in an all-male high school, the frequency of caries-free students is 10 and in a class in an all-female high school the frequency is 15, the information cannot lead to the conclusion that the frequency is higher in female school children compared to male school children and attention should be paid to the total number of students in the two classes. As an example, if there are 20 male students and 30 female students in each class, which class has a higher frequency of caries-free students? In order to compare the two classes accurately in relation to the caries-free state, it is better to calculate the frequency percentage of caries-free state using the equation below:

$$\text{Frequency percentage} = \frac{\text{Frequency}}{\text{Total number}} \times 100$$

In the example given above, the frequency percentage of caries-free students in both male and female school children is 50%.[1]

Description of Quantitative Data Using Measures of Central Tendency and Dispersion Statistics

Central and dispersion statistics are used to report research data. Central statistic refers to a value around which the majority of the data of the community cluster; they include mode, median, and mean. Dispersion statistic shows the distance of the data from the central statistic.

Measures of Central Tendency

Mode

The mode refers to the most commonly occurring value in a set of measurements. For example, in the data series below, which is again related to the length of lateral maxillary incisors in a series of 10 patients the mode is 19 because, with three cases, it is the most frequent value in the series (20.5, 19, 20.5, 19, 21, 19, 19.5, 17.5, 18, 18).

During reporting of data, it is advisable to arrange data in ascending or descending order in the beginning in order to facilitate description of data. This process in itself provides useful information about data series. For example, the data series above can be arranged in ascending order (17.5, 18, 18, 19, 19, 19, 19.5, 20.5, 20.5, 21).

Sometimes, the data series is bimodal or polymodal. For example, the data series below is bimodal: 18, 18, 19, 19, 19.5, 20, 20.5. The modes are 18 and 19.

Data series might have no modes, i.e. the frequency of values might be the same. However, such a situation is rare.

Median

If the values are arranged in ascending or descending order, the value located in the center, with half of the values higher and with the other half lower than that value, is considered as median value. For example, what is the median value in the data series below?

7, 10, 6, 11, 6, 13, 8

At first the data series should be arranged in ascending order:

6, 6, 7, 8, 10, 11, 13

Since the number of values is an odd number, the formula $\frac{n+1}{2}$ (n = the number of values) can be used to locate the median value:

$$\frac{7+1}{2} = 4$$

Therefore, the 4th value is the median.

If the number of values is an even number, the median is calculated using the mean of the two values in the middle. For example, in the data series below, which consists of the students' exam grades in statistics, the median is calculated as follows:

10, 12, 12, 13, 15, 17, 18, 18, 18, 20

$$\frac{15+17}{2} = 16$$

Mean

The most frequently used central measure in the descriptive analysis of quantitative data is mean. Mean is calculated by dividing the total sum of all the values by the total number of values based on the formula below:[3]

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

The formula shows that all the values are used to calculate the mean, which is the strong point or advantage of the mean. However, sometimes this turns out to be a weakness for the mean, especially in relation to the median. For example, if in a data series there are extreme values, the mean will be strongly influenced by them, but the median does not have such a disadvantage. Here, we once again evaluate the data series of students' grades. The mean is calculated as follows:

$$\frac{10+12+12+13+15+17+18+18+18+20}{10} = 15.3$$

Here the mean is very close to the median (it was 16).

However, if an imaginary student No.11 has not been able to study well to pass the test and has a grade of 3, the mean will decrease to 14.18. In fact, the grade of this student decreases the overall mean of class grades by 1.12 grades, which is the weakness of mean.

Generally, if extreme values or outliers exist, the median is a more appropriate statistic compared to the mean. Outlier data are those that are located very far from the mean compared to other data. In order to solve the problems of the effect of outliers on the mean, two relatively new techniques are used to calculate the mean, which include 5% trimmed and M-estimates.

5% Trimmed

In this technique the upper and lower 5% values, which usually consist of outlier data, are deleted and the remaining values are used to calculate an arithmetic mean. In large sample sizes the elimination of 10% of values does not lead to any problems; however, in small sample sizes this process has detrimental effects on data.[1]

M-estimates

In this technique, the values are given grades in terms of their importance in order to calculate the mean, i.e. the values close to the center are given higher grades and the grade decreases as the values lie farther from the center. The means of graded values are calculated. In this technique, none of the values are eliminated, but at the same time the effect of extreme values is eliminated to some extent.

Note that if the means calculated by the three techniques discussed above in a data series are almost the same, it can be concluded that no outlier values exist. If the mean calculated by the use of 5% trimmed and M-estimates techniques are less than the main mean of data, outlier values exist in the upper bounds. On the contrary, if the means calculated by the two techniques above are greater than the conventional mean, outlier values exist in the lower bound of data.[1]

Measures of Spread (dispersion)

These measures show the distance of values from the central measures. Reporting of central statistics discussed above cannot show the status of numeric series. For example, the status of dental students' grades in statistics in two 5-sample groups is as follows:

Group 1: 10, 12, 15, 18, 20

Group 2: 13, 14, 15, 16, 17

Despite the equality of the mean in both groups (15) the grades in the two groups are not similar and different variation is observed, i.e. in group 2 the majority of the grades cluster around the mean, with less dispersion.

Therefore, it can be concluded that merely reporting central measures cannot result in a good judgment about distribution of data, and it is necessary to report a measure of spread along with each measure of central tendency.

Different types of measures of dispersion include:

1. Range
2. Variance
3. Standard deviation
4. Coefficient of variation

Range

The difference between the highest and lowest values in the data series is called the range:

Range = the highest value – the lowest value

Please once again note the two series of values mentioned previously, in which the mean was 15. The range in group one was 20-10 = 10 and in group two it was 17-13 = 4.

One of the advantages of the range is that it is easily calculated. The most important problem of the range is the fact that it is only affected by the lowest and highest values in the data series. Therefore, it does not fully demonstrate the dispersion of data for the comparison of different series of data. For example, compare the two series of students' grades here:

Group 1: 10, 11, 12, 13, 14, 15, 16

Group 2: 10, 13, 13, 13, 13, 13, 16

Although the mean and range are the same in the two groups, the distributions of data are different and this is the disadvantage of the range in demonstrating the distribution of data.

Variance

Variance is a measure of spread, which is, contrary to the range, under the influence of all the data values. The following formula is used to calculate variance:

$$\sigma^2 = \frac{\sum(\bar{x} - x_i)^2}{N - 1}$$

In the formula above, in the numerator of the fraction the distance of each value (data) from the mean is calculated and then the second powers of all these values are added up, which is divided by the number of values minus 1. Therefore, it is obvious that as the distance between the values and the mean increases the variance increases and shorter distances result in a smaller variance. Now we once again consider the grades of the two student groups and calculate the variance:

Group 1: 10, 11, 12, 13, 14, 15, 16

Group 2: 10, 13, 13, 13, 13, 13, 16

As discussed previously, although the range and the mean are the same in these two groups of grades, the dispersions are different. In groups 1 and 2, based on the formula above, the variances are 4.67 and 3, respectively, indicating a greater dispersion of data in group 1. This conclusion can also be reached by one short look at the data.

Standard deviation

Standard deviation is another measure of spread and is the square root of the variance:

$$SD = \sqrt{\sigma^2}$$

Coefficient of Variation

The standard deviation and mean are influenced by the measuring technique and unit. In comparing of standard deviations of various variables care should be exercised, because if the measuring units of the variables are different, the comparison might be incorrect and misleading. For example, if the standard deviation of the length of maxillary lateral incisors is 4 mm and the standard deviation of DMF in the 6-year-old students of a school is 2, it should not be concluded that the dispersion of the root length is greater than that of caries index. In order to solve the problem, coefficient of variation is used, which is calculated by dividing the standard deviation by the mean. Therefore, due to the division carried out, the coefficient of variation does not have a unit, making it possible to carry out comparisons between variables which have different units:

$CV = \frac{SD}{\bar{x}} \times 100$

Another consideration in the evaluation of quantitative variables is distribution and its comparison with normal distribution in relation to determining appropriate statistical tests. Therefore, first, the characteristics of normal distribution will be explained.

Normal distribution

The majority of continuous variables in health sciences follow a normal distribution. This is a two parameter distribution which depends on mean and standard deviation (SD) of the variable in the population.[1] The normal distribution has a symmetric bell-shaped curve in which mean, median, and mode are the same (Figure 1).This indicates that the frequency of values around mean is much higher than tails of the curve. For example, if blood sugar (BS) level follows a normal distribution with a mean of 110 and SD of 10, the BS level of the majority of people is around 110. In addition, the BS level of half of the population is lower than 110. To be more precise, the distribution of BS values depends on both mean and SD. Around 68% and 95% of data falls between one SD and two SD around mean. Therefore, in this society, the BS level of 68% of people falls between 100 and 120. The corresponding interval that covers 95% of people is 90, and 130.

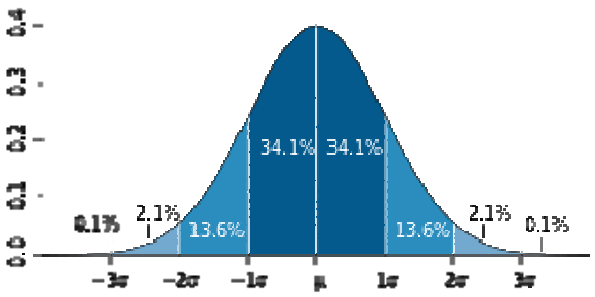

Figure 1: Normal distribution, symmetric bell-shaped curve

Dark blue is less than one standard deviation away from the mean. For the normal distribution, this accounts for about 68% of the set, while two standard deviations from the mean (medium and dark blue) account for about 95%, and three standard deviations (light, medium, and dark blue) account for about 99.7%

Normality is the underlying assumption behind most statistical techniques. Therefore, the investigation of this fundamental assumption, before planning for data analysis, is important. There are a variety of numerical, graphical, and P-value based tools to check whether data merit this assumption.

We should emphasise that investigation of normality necessitates the careful exploration of data. Each of the methods noted have their own advantages and disadvantages. Here we only present two descriptive statistics frequently used in the literature, known as skewness and kurtosis.

We noted that normal distribution has a symmetric curve. Skewness refers to a lack of symmetry. In the normal distribution we expect 5% of data to be out range of mean plus/ difference two SD. Data are skewed when the proportion of values in tails is contrary to our expectation from normal distribution. Skewed distributions have a tail in right (known as right or positive skewed) or left (left or negative skewed) (Figure 2). To estimate skewness one can simply calculate the difference between mean and median, and divide it by SD. Values between -1 to 1 indicate that the normality assumption is reasonable.[4]

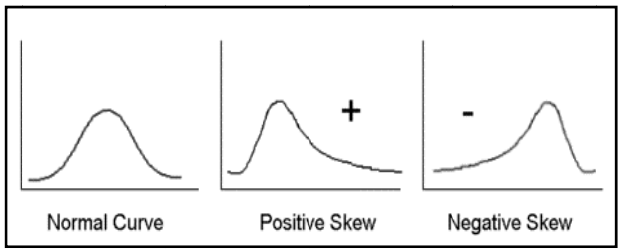

Figure 2: Positive and negative skewness vs. normal distribution

Kurtosis is another statistics which indicates whether the data are peaked or flat relative to a normal distribution. That is, data sets with high kurtosis tend to have a distinct peak near the mean, decline rather rapidly, and have heavy tails. Data sets with low kurtosis tend to have a flat top near the mean rather than a sharp peak (Figure 3). Similar to skewness, values between -1 and 1 justify the normality assumption.

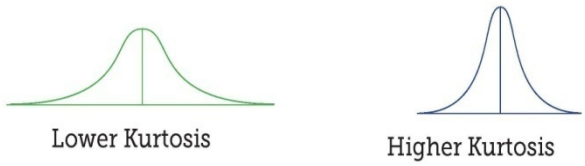

Figure 3: Lower kurtosis distribution vs. higher kurtosis

Descriptive statistics for qualitative variables in SPSS

In working with qualitative variables, the statistics to be reported are frequency, and percentage. You should select analyze, descriptive statistics, and frequencies. Then select categorical variables from the left box and transfer them to the right box. Finally, simply press the OK button. In the output window, the first table provides information about the percentage of missing and valid data for each variable. Then, the next table gives frequency, percentage, valid percentage, and cumulative percentage. We should note that in calculation of percentage the denominator is the total sample size, while in valid percentage the denominator is the number of subjects with available data.[1]

Descriptive statistics for quantitative variables in SPSS

Relevant statistics for continuous variables involve mean, median, percentiles, SD, variance, range, and etc. SPSS provides multiple tools for these statistics. These tools can be used from the frequency, descriptive, and explore menus. Details are given below.

Frequencies menu:

We have explained the use of this menu for qualitative variables. However, some descriptive statistics for quantitative variables can also be calculated through this menu. You must select analyze, descriptive statistics, and then frequencies. Remember to deselect the display frequency table. Otherwise, you will find a long frequency table for a quantitative variable in the output window, which is useless. Transfer quantitative variables from the left to the right box, and then press the Statistics button. This opens a new window in which you can select the central tendency statistics (such as mean, median, and mode), and dispersion statistics (such as min, max, variance, SD, and SE). In addition, if you select the quartiles in the percentile values box, the software will provide the first, second, and third quartiles. If you select percentile, you can ask the software to provide any percentile you wish. Finally, in the distribution box, you can select skewness and kurtosis statistics to check whether data follow a normal assumption or not.

Descriptive menu:

Again you must select analyze, descriptive statistics, and descriptive. Then, transfer the quantitative variables to the variable box and click the options button. From the central tendency statistics you can only select mean. This path does not provide median or mode. In addition, no option for percentiles is available. Other statistics can be selected similar to the frequency approach.

Explore menu:

Here select analyze, descriptive statistics, and explore. Transfer quantitative variables to the dependent list box. This path provides the opportunity to provide statistics across levels of qualitative variables as well. To do so you should select qualitative variables from the factor list box. If you do not select any variable from this box, the overall statistics will be calculated. In the output window, you will automatically find mean, 5% trimmed mean, confidence interval of mean, and etc. By selection of the statistics button, you can select some new statistics such as m-estimators of mean, and outliers.[1] In addition, it is possible to get graphs such as box plot and steam and leaf plot.

The next paper in this series will take a further look at using tables and graphs in dental articles.

Conflict of Interest

Authors have no conflict of interest.

References

1. Chehrei A, Haghdoost AA, Fereshtehnejad SM, Bayat A. Statistical methods in medical science researches using SPSS software. 1st ed. Tehran, Iran: Pazhvak Elm Arya; 2010.
2. Petrie A, Bulman JS, Osborn JF. Further statistics in dentistry: Part 1: Research designs 1. Br Dent J 2002; 193(7): 377-80.
3. Von Fraunhofer JA. Research writing in dentistry. New York, NY: John Wiley & Sons; 2009.
4. Chan YH. Biostatistics 101: data presentation. Singapore Med J 2003; 44(6): 280-5.