Received: 18 Aug. 2015 **Accepted: 16 Nov. 2015**

Cleansing and preparation of data for statistical analysis: A step necessary in oral health sciences research

Hossein Molavi Vardanjani MSc, PhD[1], *Ali Akbar Haghdoost MD, MSc, PhD*[2], *Arash Shahravan DDS, MSc, PhD*[3], *Maryam Rad PhD*[4]

Review Article

Abstract

In many published articles, there is still no mention of quality control processes, which might be an indication of the insufficient importance the researchers attach to undertaking or reporting such processes. However, quality control of data is one of the most important steps in research projects. Lack of sufficient attention to quality control of data might have a detrimental effect on the results of research studies. Therefore, directing the attention of researchers to quality control of data is considered a step necessary to promote the quality of research studies and reports. We have made an attempt to define the processes of cleansing and preparing data and determine its position in research protocols. An algorithm was presented for cleansing and preparing data. Then, the most important potential errors in data were introduced by giving some examples, and their effects on the results of studies were demonstrated. We made attempts to introduce the most important reasons behind errors of different natures; the techniques used to identify them and the techniques used to prevent or rectify them. Subsequently, the procedures used to prepare the data were dealt with. In this section, techniques were introduced which are used to manage the relationships established between the premises of statistical models before carrying out analyses. Considering the widespread use of statistical models with the premise of normality, such premises were focused on. Techniques used to identify lack of normal distribution of data and methods used to manage them were presented. Cleansing and preparation of data can have a significant effect on promotion of quality and accuracy of the results of research studies. It is incumbent on researchers to recognize techniques used to identify, reasons for occurrence, methods to prevent or rectify different kinds of errors in data, learn appropriate techniques in this context and mention them in study reports.

KEYWORDS: Cleaning; Preparation; Statistics; Data; Quality Control

Citation: Molavi Vardanjani H, Haghdoost AA, Shahravan A, Rad M. **Cleansing and preparation of data for statistical analysis: A step necessary in oral health sciences research.** J Oral Health Oral Epidemiol 2016; 5(4): 171-85.

Research in the field of health, like other fields, consists of structured efforts to answer a question or solve a problem. The procedural steps of a study consist of designing, making sure of the quality of study procedures, implementation of procedures, collection, quality control and analysis of data, and finally reporting the results of the study.

It is obvious that if a study is designed and implemented in a more academic and more accurate manner, the probability of finding appropriate responses for research questions will increase. Therefore, the quality control of data in all the stages of the study is very important. It is possible to define critical points for quality control of data in the research process, which might include the

1- Assistant Professor, Department of MPH, School of Medicine, Shiraz University of Medical Sciences, Shiraz, Iran
2- Professor, Research Center for Modeling in Health, Institute of Futures Studies in Health, Kerman University of Medical Sciences, Kerman, Iran
3- Professor, Endodontology Research Center AND Oral and Dental Diseases Research Center AND Kerman Social Determinants on Oral Health Research Center, Kerman university of Medical Sciences, Kerman, Iran
4- Assistant Professor, Oral and Dental Diseases Research Center, Kerman university of Medical Sciences, Kerman, Iran
Correspondence to: Arash Shahravan DDS, MSc, PhD
Email: a.shahravan@kmu.ac.ir

procedures of data collection, classification, coding and entry into a software program.[1] If the researcher does not make an effort to control quality of data at each of the critical points, it is possible that the accuracy of the study results will be severity compromised. However, despite the efforts made by researchers to control the quality of data, it is possible that human errors, especially in multicenter and national studies, will not be completely eliminated.[2]

Quality control of data at each critical point consists of efforts to identify errors and their types and to determine the best technique to deal with them, considering the existing condition. Let's give easy examples to review the most important potential errors in epidemiologic studies:

Example A

Suppose that a researcher in trying to answer the following questions: "What is the mean serum cholesterol level of elementary school students (under 10 years of age) and how many hours do they watch TV?" To answer these questions, the researcher should take a blood sample from all the students using standard tools and record the number of hours each student watches TV. If for some reason or another the researcher is forced to first take blood samples from the students and then determine the number of hours they watch TV, what will happen? You might have guessed the answer correctly; the number of hours some students watch TV cannot be determined and they will be "missing."

Example B

Now suppose that the researcher is entering the serum cholesterol levels of the students in example A, which have been recorded on special laboratory sheets, into a software program. The researcher unknowingly types the last digit of the serum cholesterol level twice (1100 instead of 110). If we suppose that the normal range of serum cholesterol levels is a circle, it can be imagined that the value 1100 has been "thrown" out of the circle.

Example C

Suppose that a researcher wants to estimate the incidence of different cancers separately in Iran using the data of the National Data Center for Registering Cancer Cases. To this end, the researcher needs to list the cancer cases separately in relation to tumor location, age, and gender. After preparing the list, the researcher realizes that the gender of some prostate cancer patients has inadvertently been recorded as female. The researcher knows that female gender and prostate cancer are not "consistent."

In the examples above, three types of errors in data were explained. A large number of examples can be given in which each type of such errors occurs due to a reason other than the above. The technique used to deal with each error is dependent, to some extent, on its cause. Therefore, it is necessary for researchers to be acquainted with different kinds of errors in data, their etiologic agents and the appropriate techniques to tackle them. A set of activities carried out to identify errors, the type of error and the possible etiologic agent and finally the technique used to tackle the errors found in data is referred to as "data cleaning."[3] Cleansing of data is one of the most necessary steps after collection of data and before analysis of data.[4] In the sections to follow, we have made efforts to review the general framework of this vital step. To being, figure 1 presents the process and the steps to cleanse and prepare data for statistical analysis, followed by the review of each step in the text.

Inconsistent data

We direct our attention to the example "C" above, in which during registration of data, a female patient had been registered as having prostate cancer. In this example, based on definition of gender, having prostate cancer is inconsistent with being a female. In general, the inconsistency of data is divided into two groups: definition-based and data distribution-based.[5]

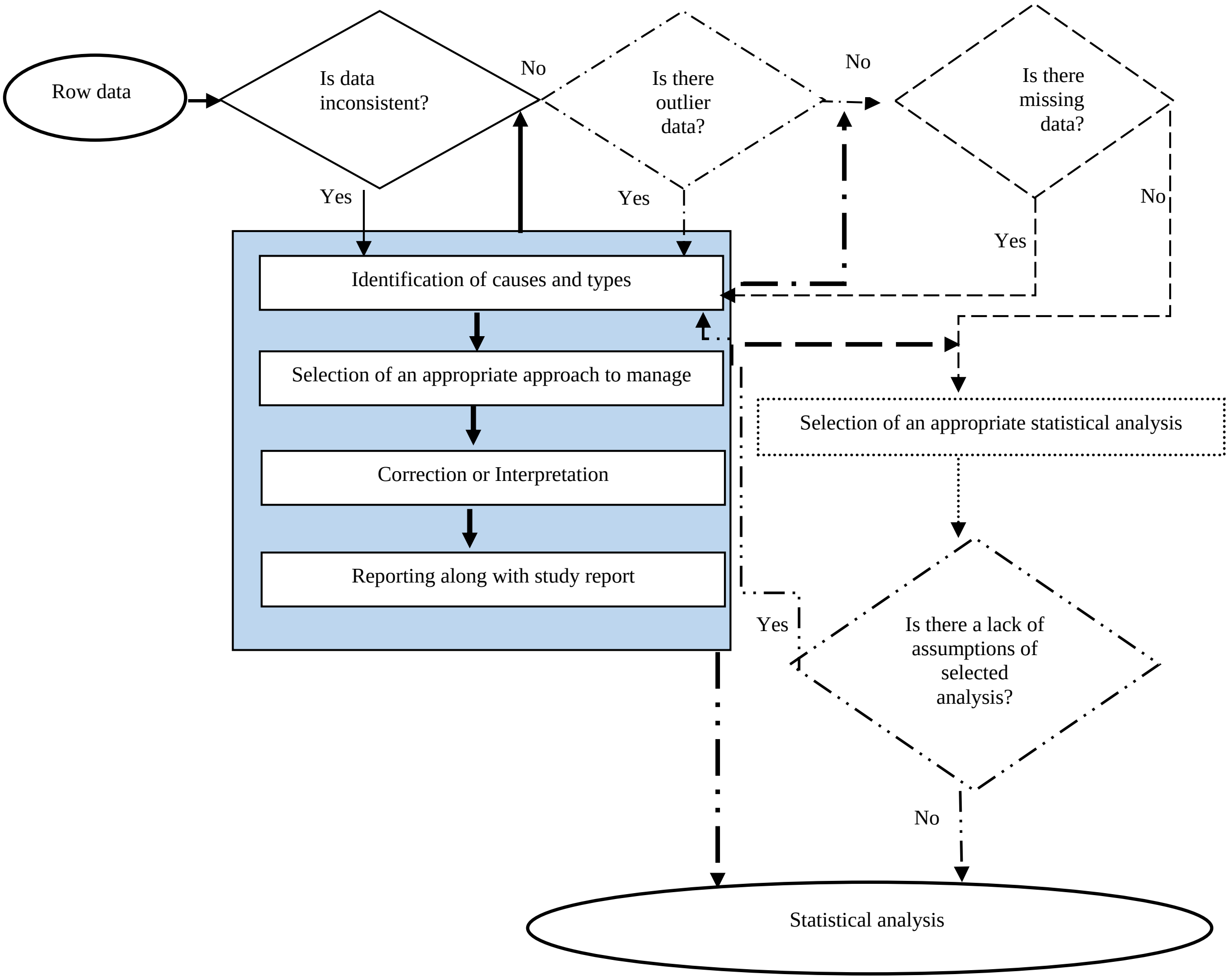
Row data
Is data inconsistent?
No
Yes
Is there outlier data?
No
Yes
Is there missing data?
Yes
No
Identification of causes and types
Selection of an appropriate approach to manage
Correction or Interpretation
Reporting along with study report
Selection of an appropriate statistical analysis
Is there a lack of assumptions of selected analysis?
Yes
No
Statistical analysis

Figure 1. The process of cleansing and preparing data on health-related research

Therefore, outlier data, too, can be considered inconsistent data, which do not necessarily happen due to error. However, it is certain that inconsistency, based on definition of data, is due to errors. This type of error can cast doubts on the validity of study results. Therefore, it is necessary to evaluate the consistency of an individual's data with his/her other data and an individual's data with data from other individuals.

To evaluate the consistency of data, it is necessary to define and determine criteria for consistency of data before data collection and after entering data into the software program, consistency of data should be evaluated using the criteria defined. Such evaluation can be carried out with the use of crosstabs with qualitative variables. Drawing of such tables is very easy in many software programs, such as the following path in SPSS (SPSS Inc., Chicago, IL, USA):

SPSS → Descriptive Statistics → Crosstabs

In relation to quantitative variables, techniques can be used that are introduced in the section on outlier data and also on comparison of data with the possible minimum and maximum values for the variable.

SPSS → Descriptive Statistics → Descriptive

Errors during entering data into software programs

As disused previously, one of the critical points for the occurrence of errors in data is when the data are entered into a software program. The errors occurring at this stage can easily be identified and corrected. It is necessary to note that entering data into software programs does not always involve entering data from forms or questionnaires into a software program. Sometimes, transfer of data between different software programs or combining data files can result in errors, especially in inconsistency between them. For example, suppose we have two files. In the first file, the gender has been defined as female = 1 and male = 2 and in the second file as female = 2 and male = 1. Although both files have undergone quality control before being combined, if the quality control of data is not carried out after they have been combined it is possible that all the previous efforts for the quality control of files will be compromised because finally data containing contaminated information will be used. Therefore, it is suggested that before combining or making any changes in data reassures be designed and implemented to prevent such errors.

Techniques used to prevent errors at the time of entering data into software programs are divided into two major groups.

1. Visual control of data: In this technique, the researcher compares the data entered with the data written on paper or data in the original file after entering all the data into the software program and corrects all the inconsistencies. A modified version of this technique is reading out data by one researcher and matching of data entered into the software program by another researcher.

2. Re-entering data: In this technique, the researcher enters the data twice into two separate files and then compares them and corrects the inconsistencies. Sometimes, a modified version of this technique is used, in which two researchers enter the data separately into the software program and then comparisons and corrections are made.

Both of the techniques above can initially be carried out for only a percentage of data and then decisions can be made to continue or stop the process.

Outlier data

In the example B above, 1100 mg/dl for serum cholesterol level is considered outlier data. Imagine that in the example B only one outlier data exists (1100 mg/dl = the serum cholesterol level of student X), and there are 100 students in the study, whose mean cholesterol serum level (n-1 = 99) is 120 mg/dl. With only this outlier data, the mean will increase from 120 mg/dl to approximately 130 mg/dl. It can be concluded that if similar cases of outlier data exist in this example the mean cholesterol levels will dramatically be overestimated.

If outlier data are not identified and corrected, they can influence the distribution of data, exerting detrimental effects on the results of the study. In summary, outlier data might increase variance and usually decrease the statistical power of analyses, alter the potential type I and type II statistical errors, decrease the normal distribution of data (if they occur in a non-random manner) and lead to bias in estimation of statistics.[6]

It should be pointed out that not all the outlier data occur due to errors.[7] In relation to the etiologic factors, outlier data can be classified into four groups as follows: due to errors in location, time and measuring technique and instrument, in reporting by the participants, in registering or in entering into the software program; due to sampling from different populations (If data of one population/group is erroneously entered into the data of another population/group, the data are called a “contaminant”); due to the skewness of the variable being measured (higher than the expected distribution for it); due to relatively rare occurrences(what mainly lead to influential observations).[8] After identification of outlier data, it is necessary to carry out further evaluations to identify their potential causes and make a proper decision about each outlier data based on these evaluations.

Various suggestions have been made in relation to the definition, identification and dealing with outlier data,[7] of which one useful suggestion is to classify these data into two groups of univariate and multivariate.

Univariate outlier data

In cases in which the amount of one variable for one subject is very different from the amount of the same variable for other subjects, that value is considered an outlier data (Example B).[4] One of the most common criteria for identification of such data in quantitative variables is to define ± 3 standard deviation (SD) from the mean value as a normal or expected range for the data (This value in the Schweinle method has been defined as 2.5-folds.). Based on this criterion, if one data is beyond the ± 3 SD range, it will be identified as an outlier data (This method will not be appropriate when the sample size is very small and the expected distribution of data is very different from the normal distribution). In other words, if the Z-score (Z-score is calculated by subtracting the numeric value of one variable from the mean of the same variable and the result is divided by the SD) of one data is > 3, it is considered outlier, requiring further evaluation. Why data beyond this range is considered outlier can be explained by the fact that if the distribution is normal, there is only a 0.26% probability that a data would be placed beyond this range. This percentage is a very low probability, and therefore, it is logical that the validity, mechanism of creation and accuracy of this data will be dubious.[9]

In addition to the calculation of the Z-score statistic, there are different other techniques, too, to identify outlier data. The most commonly used techniques are Grubbs' test, Dixon's Q-test, variance graphs, box histograms, and use of interquartile range.[10,11] The reader can refer to the relevant references for further details. Based on your choice for identification of outlier data, different pathways can be suggested in each software program. One of the easiest techniques in SPSS can be shown as follows:

SPSS → Analyze → Descriptive Statistics → Explore → Statistics → Outliers

Using this pathway, it is possible to request calculation of one of the robust methods to estimate a mean referred to as 5% trimmed mean.

Multivariate outlier data

A multivariate outlier data are a data which is not logical by considering several variables simultaneously.[4] For example, data of an individual with a height of 190 cm but a weight of 45 kg does not seem logical. Imagine the fact that in this example, having a height of 190 cm or a weight of 45 kg are not outlier data when considered separately. Leverage and discrepancy parameters are used to identify multivariate outlier data.

Leverage parameter shows how distant each observation is from other observations. However, this parameter cannot show whether an observation that is far from other observations is on the same track of other observations or not. In fact, the leverage parameter does not provide information about the direction of distance of one observation from other observations. To obtain this information, discrepancy parameter is used. Figure 2 shows how the two leverage and discrepancy parameters show the distance of one outlier data from other data.

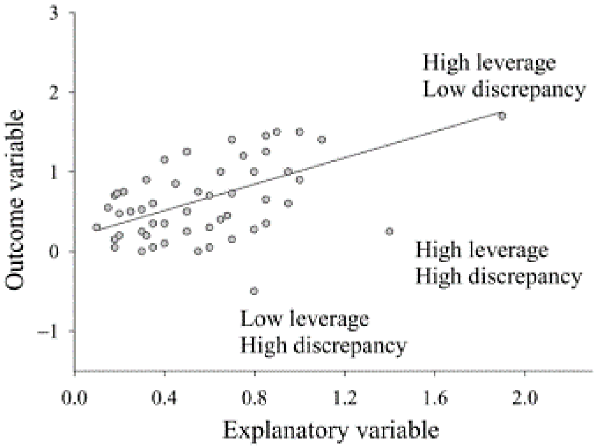

Figure 2. Leverage and discrepancy parameters for determining outlier data

Cook's distance (Cook's D) is a parameter which gives us a combination of leverage and discrepancy data. It tells us how much the regression coefficients will undergo changes if one observation is eliminated. To identify influential outlier data through Cook's distance, the 4/(n−k−1) ratio is used, where "n" is the sample size and "k" is the number of independent variables.

How should outlier data be dealt with in your idea? As discussed above, it is necessary to carry out further evaluations to determine their etiologic factors and then take measures to correct, interpret or eliminate them. To this end, it is necessary to first check the accuracy of data by evaluating whether errors have occurred during measurements or during recording of data and whether errors have occurred during entering of data into the software program. If the answers to these questions are positive, it is obvious that it will be necessary to correct data. In the next stage, the following question is asked: "Has a subject out of the target population undergone a sampling procedure during sampling?" If the answer is "yes" one of the most acceptable techniques is to eliminate the outlier data from the analyses and pay attention to it at the data analysis stage.[5]

At this stage, the number of outlier data that cannot be corrected is of utmost importance. If the number of such data is insignificant relative to the sample size, e.g., one or two outlier data in a database with more than 100 samples, no significant changes will be observed in statistics and results. However, if there are multiple outlier data or if the sample size is small, other techniques should be used to mitigate the effects of outlier data, which include transformation of bases, truncation of distribution and use of robust methods.[12]

Missing data

Let's return to example A, where despite the attempts made by the researcher, it is not possible to collect data in relation to the number of hours some children watch TV. Statisticians call these data "missing data." It's no exaggeration if we claim that almost none of the quantitative data will be free of missing data (222-2-1801-2193). Missing data can result in serious bias in the results or a decrease in the power of statistical analyses.[13,14] The magnitude of such an effect depends on various factors, including the pattern of missing data and the reason for or mechanisms of their being missed and to a lower degree on the percentage of missing data.[15]

In general, missing data can be classified into three groups in relation to the mechanism of their occurrence. The "completely random missing" data occur when there is no regular difference between missing data and the observed data; in other words, the odds of being missed are equal for all the values of a parameter for all the participants in the study. Ignoring such missing data only results in a decrease in the statistical power of analyses, without any biases in the results. The premise of "completely random missing" data can only hold true for a limited number of cases. Lower levels of randomization can be found in "randomly missing" data.

The basis for defining the randomly missing data group is the fact that there are regular differences between the missing data and the observed data, but these differences can only be explained with the use of other variables which have been measured in the study. Remember example A. If we register the gender and age of all the students and can assume that missing of data on watching TV is only due to differences in the students' gender and age, the missing data in this example can be considered random. Certainly, the premise that all the missing data are random in all the studies is not rational, either. These data can be replaced with the use of a wide range of statistical techniques.[16,17]

In some studies, missing of data is dependent on variables which either have not been measured in the study or the values or measurements of some participants have been lost. This kind of missing data are referred to as "missing not at random" data. This type of error can significantly decrease the internal and external validity of the study results. A well-known example is the attrition of participants in a clinical trial due to the side effects of the treatment modality used (in cases in which the side effects are not measured or evaluated). Different functional reasons lead to the missing of data, including a lack of response by participants to some specific questions, collecting data at more than one episode, incorrect sequence of measuring different variables, lack of familiarity of the researcher or interviewer with the techniques used to encourage cooperation and establish a constructive relationship, inappropriate conditions of the measurement environment, inattention to the cultural conditions and considerations of the participants, laziness, negligence and forgetfulness of the questioner or the individual in change of recording data, attrition of participants and errors during entering data into the software program.

Although it was explained above that one of the considerations in relation to the evaluation of the effect of missing data is the percentage of these data, it is less important than the mechanism and the pattern of missing of these data.[18] On the other hand, although some statisticians have suggested that 5-10% is the maximum acceptable level for such data, no accepted critical level has been defined for the percentage of missing data.[19]

In discussions on the pattern of missing data, it is possible to define three different patterns. Imagine a study in which the numerical values of variable K (V_1, V2,…, V_k) are measured: (1) If the values of one or more variables of some participants (e.g., V_2, V_1 or V_k) are lost, the missing data pattern is referred to as univariate. (2) Now suppose that values of V_5 are lost and as a result the data of V_6 to V_k are lost, too. This is due to the dependence of variables on one another or their time sequence. To understand this better, suppose that in a clinical trial with repeated measurements, V_5 is the fifth measurement, which is lost due to participant attribution. In that clinical study, it is highly probable that subsequent measurements will be lost due to the continuation of lack of cooperation. Therefore, in such situations, data will be lost from one point on. This pattern is referred to as homogeneous pattern. (3) In the third pattern, referred to as irregular pattern, values for each variable of some participants are lost in a random manner.

The best critical point to deal with missing data is during collection of data. In other words, prevention is always better than correction of this error. In addition, correction of such a serious deficiency in data is helpful and even necessary by replacing the missing data or values. Of course, before correcting such deficiencies, it is necessary to identify the details of the deficiency by analyzing the missing data.[3] There are various techniques to analyze missing data and replace them, which are beyond the scope of this article. A very important consideration is the fact that use of each technique and the validity of their results depend on various considerations and factors, such as mechanism, pattern, percentage of missing data, and sample size.

Management of missing data in statistical software programs, too, is of great significance, which can be useful in identification and proper management of such errors. Different statistical software programs use different signs to show missing data. One of the most important signs in this respect is point [.]. Use of a point with a numeric definition is better than the numeric symbols of 9 or 999 because if these numbers are not defined as codes for missing data, they might be considered real data during analyses and influence the results of the study.

In the majority of statistical software programs, special commands and menus have been designed for the analysis and replacement of missing data (due to the wide range of these items, they will not be discussed in detail here). In the early stages of reviewing data, it is possible to use the frequency command in different software programs (Analyze → Descriptive Statistics → Frequencies in SPSS). This command can determine the number of missing data for each variable.

Observations with various error types

In some cases, more than one error type mentioned above occurs in data in one observation (e.g., in data of one participant). In such cases, if the errors in the data still remain after referring to the registered documents of the observation in question (e.g., the paper version of the questionnaire of that participant), the best recommendation is to eliminate that observation. Therefore, if one or several variables of one observation are outlier data and some others are missing data, the best option is to refer to the paper documents of that observation. If the errors are not corrected, in the next stage, this observation will be eliminated and will not be included in the analyses of missing data.

Establishment of premises for the analysis of data

After the researcher evaluates the quality of data in the previous stages and corrects the deficiencies, it is necessary to select an appropriate statistical model for the analysis of data. In this context, there exist many important considerations. Here, it is necessary to note that the majority of statistical analyses have some premises and correct results will be obtained on the condition that they are established.[20] Another important consideration in the selection of an appropriate model for analysis of data is the fact that the researcher should select the simplest appropriate statistical model. Although discussions on the steps involved in selecting an appropriate statistical model is beyond the scope of this article, attempts have been made to review one of the most common steps in the following sections.[21]

Normality of distribution of data

As discussed previously, it is necessary to control the establishment of premises of the selected model for analysis of data. The premise of a relatively large group of simple and commonly used models for the analysis of quantitative data is normality of distribution of data.[20] Normality means that the histogram of data is similar to the normal distribution graph. There are different techniques to evaluate normality of data. A very important consideration is the fact that none of the techniques should be used as the only criterion for judgment. A correct decision for the normality of data is possible based on the results of several appropriate methods. In the following paragraphs, some of these methods will be explained.

Comparison of central parameters of data

In the normal distribution of data, the numerical mean, mode and median are equal and the same. Therefore, comparison of these three parameters can provide information about the normality of data distribution. For example, as the absolute value of the parameter D becomes smaller the distribution of data gets closer to normal distribution.

$$D = \frac{\text{Median} - \text{Mean}}{\text{Mean}} \times 100$$

Use of data distribution parameters and the properties of normal distribution

The range of changes of the variable can provide information about the normality of data distribution. To use the range of changes, one of the properties of normal distribution is used (in normal distribution, 95% of data lie within 2 SDs from the mean). To this end, the value "2 × mean ± SD" is calculated for the variable under question and compared with the range. If the difference is low (approximately 5%), the distribution is probably normal.[4]

The skewness and kurtosis parameters show the inclination of data to one-tail (a measure of horizontal symmetry of distribution). Skewness can be used in two-ways to make a decision about the normality of data distribution. One method is to make a decision about the severity and direction of skewness of "sample distribution" based on parameter value. There are different methods to quantify skewness, but the aim of this article is not to explain these items. Therefore, suffice it to say that the amount of skewness of an ideal normal distribution is zero and for non-normal distributions it is a positive value (skew to the right) or a negative value (skew to the left). Therefore, skewness values from −1 to +1> mean severe skewness; values from −1 to −1.2 and from +1 to +1.2 mean moderate skewness and values from −1.2 to +1.2 indicate relative skewness.[22]

Another method is to calculate the skewness standard statistic values and carry out the statistical test to compare it with the critical value from −1.96 to +1.96 at a confidence interval (CI) of 95%. This technique will provide us with some information about the normality or non-normality of the "distribution of the variable in the population." It can be understood that the meaning of this awareness about what determines the amount of skewness is very different. To calculate the standard statistic of skewness, the parameter value is divided by the standard error. If the statistic is placed at a range from −1.96 to +1.96, the distribution of variable is not skewed; otherwise, the hypothesis of "not skewed" will be ruled out.[4]

If the hypothesis of "not skewed" is confirmed, it is necessary to assess the kurtosis of data distribution, too. Kurtosis shows the peakedness of the distribution summit in comparison to the summit of normal distribution. As discussed above in relation to skewness, in making decisions about the normality of distribution of data, it is possible to make two uses of the kurtosis parameter. If data are distributed normally, the numeric value of kurtosis will be 3. However, in some software programs (such as SPSS and SAS), the parameter value is not displayed and only the numeric value of the difference of distribution kurtosis from 3 is displayed, which might be a positive or a negative value. To use the kurtosis parameter, the "index value" is compared with 3 or the "numeric value of the difference of the parameter from 3" is compared with zero. If the parameter value is < 3, the distribution peak is shorter than the normal distribution peak (platykurtic), and vice versa (leptokurtic) and if it is equal to 3, the distribution peak height is equal to that of normal distribution (mesokurtic). The second use of the kurtosis parameter is the possibility of carrying out the statistical test with a premise that the kurtosis parameter is equal to 3. To achieve this aim, the procedures will be the same as those discussed above about skewness.

If the skewness and kurtosis of the distribution of the variable in question are similar to normal distribution, the premise of the normality of data distribution is confirmed; otherwise, the premise is ruled out. It should be noted that we have to consider other methods when we are going to decide on normality.[21]

To use any of the methods mentioned above, the descriptive statistics of each of the statistical software programs can be used to acquire the necessary information. For example, in the SPSS, it is possible to place the variable in questions on the dependent list in the following pathway:

Analyze → Descriptive Statistics → Explore

And calculate the parameter which is required.

Statistical tests

Although these tests alone cannot provide a definite response in relation to the normality of data distribution, if they are used correctly, they can provide more accurate responses compared to previous methods. Various statistical tests are available to evaluate normality of data distribution, the most commonly used of which are Shapiro-Wilk (S–W) and Kolmogorov–Smirnov (K–S) D test (Lilliefors test).

Each of these tests has specific characteristics, and they should be used based on the characteristics of data in question.

For example, S-W test will have a proper performance with a sample size of 7-2000. With larger sample sizes, other tests such as

Shapiro–Francia, Skewness–Kurtosis, and Jarque–Bera can be used. The K–S test has lower sensitivity to a lack of normal distribution of data compared to the S–W test.[22]

It is not possible to apply all the available tests for the evaluation of normality in any of the statistical software programs. However, it is possible to run the two commonly used tests of K–S and S–W with SPSS. The following pathway is available in SPSS for these two tests:

SPSS → Analyze → Descriptive Statistics → Explore →Plots → Normality plots with tests

To use the pathway above, it is necessary to select the lowest box of the explore window in one of the options of plots or both. The null hypothesis in both these tests is the normality of data distribution; therefore, statistical significance level of < 0.05 indicates the difference in the distribution of the variable in question relative to the normal distribution. If the results of both these tests show the normal distribution of one variable, the parametric statistical methods can comfortably be used for that variable. However, if the test results are significant and it appears that the distribution of data is not normal, it cannot reliably be claimed that non-parametric statistical methods should be used and this consideration should be kept in mind when the sample size is large. Another important consideration is the fact that these tests do not provide any information about the reason for non-normal distribution of data; therefore, attention should always be paid to other methods used to evaluate the normality of data, such as skewness and kurtosis variables and the graphs used to assess the normality of data, which will be discussed in the following section.

Graph-based methods

In one classification system, the methods used to assess normality of data are divided into two groups: numerical and graphical. Methods presented up to this point are all considered numerical methods. In the graphical methods, decisions are made based on comparison between different graphs with one definite and standard pattern. In this group, different types of graphs, including stem-and-leaf plot, detrended normal quantile-quantile (Q-Q) plot, Q-Q plot, histograms and (skeletal) box plots exist. In the following sections, each graph will be introduced. It should be pointed out that each graph contains different data and should be assessed differently.

Histogram

This graph provides visual data on the nature of distribution and its similarity to the bell-shaped graph in the normal distribution based on the frequency of drawn observations. If a gap exists in data in this graph or the distribution of data has more than one mode, it will be identified. In addition, outlier data and their distance from other data can somehow be identified by looking at the graph. Figure 3 (A and B) is examples of histograms with normal and non-normal distributions, respectively.

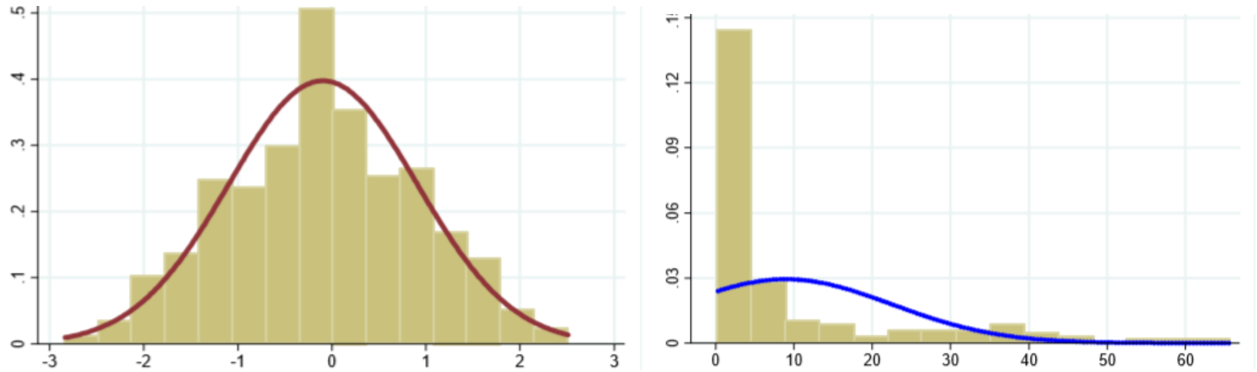

Figure 3. A histogram with normal distribution (A) and histogram with non-normal distribution (B)

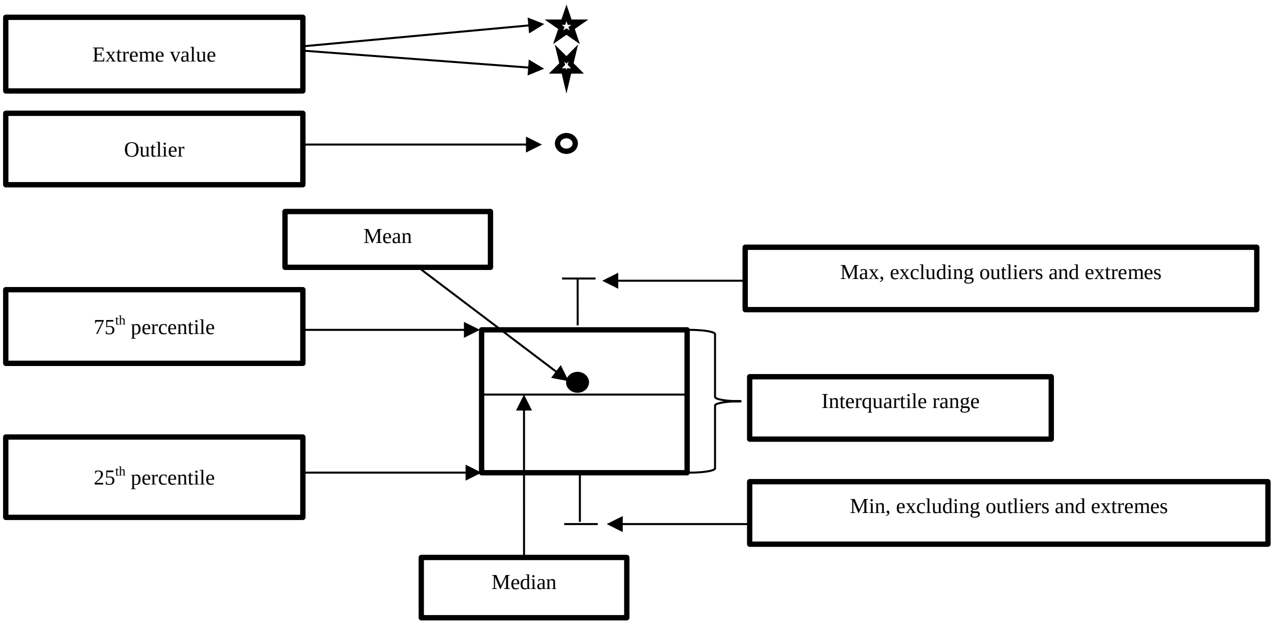

Figure 4. Schematic representation of a box plot (O = Outlier data; * = Extreme value)

Stem-and-leaf plot

In cases in which data can be shown using integers, this graph can replace a histogram to show distribution of data. There are many techniques and details for drawing graphs; however, in general, in such a graph there exists a vertical line. The stems and leaves are located on the right and left sides of the vertical line, respectively. With an increase in the resemblance of leaf section to a bell, there is an increase in closeness of distribution to normal distribution.

Box plot

This graph consists of a box and two whiskers around the box. The horizontal line drawn at the middle of the box indicates the median of data and its two parallel sides indicate the first and the third quarters; therefore, the height of the box equals the interquartile range. The point which has been determined within the box indicates the mean of data. The whiskers show the minimum and maximum of data at "1.5 × distance" from the interquartile range. In this graph, the distance of the first quarter from the third quarter minus "1.5 × interquartile range" is called the inner fence and the distance of "3 × interquartile range" from the first and third quarter is called the external fence. If each data is located between the outer fence and the inner bound, it will be referred to as the outlier data and if is located beyond the outer fence it will be referred to as an extreme value (Figure 4).

If there are multiple outlier data or extreme values or if the median line is not located in the middle of the box (corresponding with the mean), there are doubts that the distribution is non-normal. If the horizontal line approaches the upper or lower sides, the data are positively or negatively skewed, respectively.

Normal Q-Q plot

In normal Q-Q plot, the real position of each data is drawn relative to the position the data should have if the distribution is normal. To show this pattern, the plot has two sections: a sloped line which shows the ideal conditions of normality (when the observed value and the expected value are equal on the condition of normality) and hollow circles which indicate the real observed value compared to the expected value on the condition of normality. If the distribution of data is normal, the points will almost be located on the sloped line. Any deviation from the sloped line means deviation from the normal distribution. As the total or collected distance of the points increases from the sloped line, there is more deviation from normal distribution.

It should be noted that this plot does not provide any information about outlier data. If distribution of points begins from somewhere above the sloped line, moves to the underneath of the line and then ends above the line, it indicates skewness to the left. If the crescent-shaped pattern above is reversed, ii indicates skewness to the right. If the distribution pattern resembles a long S, with data variance pattern of above-under-above-under, it indicates distribution with positive kurtosis and if the pattern is reversed, i.e., under-above-under-above, it indicates negative kurtosis. It is possible in this S-shaped pattern for only a part of the central segment of S to be placed on the other side of the sloped line, in which apart from the non-normal kurtosis, distribution of data is skewed, too (Figure 5).

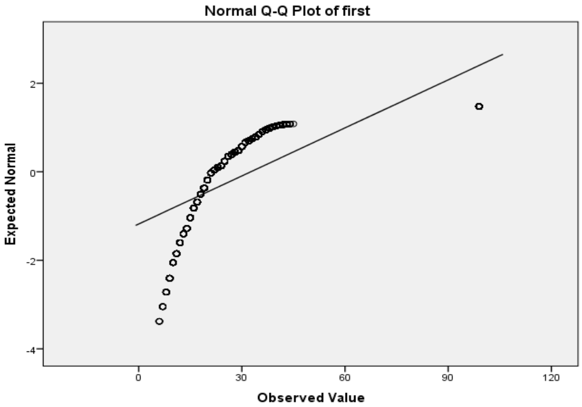

Figure 5. Normal quantile-quantile (Q-Q) plot of a non-normal distribution

Detrended normal Q-Q plot

In this plot, the amount of deviation of point observations from the direct and horizontal line is displayed as a sign of normal distribution. If the distribution is normal, the point observations will be distributed randomly and evenly above and under the horizontal line. However, if the distribution is not normal, the points will be placed in the form of reverse J or U letters and the line of normal distribution is not located in the middle of data (Figure 6).

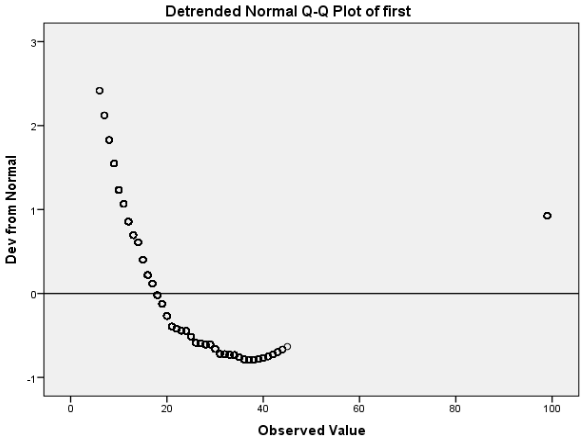

Figure 6. Detrended normal quantile-quantile (Q-Q) plot in the form of letter J, indicating non-normal distribution

As discussed at the beginning of this section, none of the methods above alone can make us sure that distribution of data is normal or non-normal. Now imagine, based on the results of all these methods, we reach the conclusion that distribution of data is a little different from normal distribution. Does any deviation mean that commonly used and accepted statistical tests (parametric tests) cannot be used? In the next section, this question is going to be reviewed.

Transformation

Although a lack of establishment of premises of a statistical method is very important and provides a definitive reason for not using that method, regarding the high rate of favorability and ease of interpretation of commonly used parametric methods and the limited diversity and low statistical power (almost 5% less compared to corresponding methods) of non-parametric methods, statisticians have suggested methods to compensate compromise of the premises of common and parametric tests. Therefore, it is possible to compensate minor deviations from normality hypothesis with transformation methods and if the transformation proves useful, it is possible to analyze data with parametric statistical tests.[4]

To this end, if data distribution is skewed positively, transformations of the square root, logarithms at a base of 10 or Napier number (e) and reversing can be used. If data distribution is skewed negatively, the value should first be reflected and then again one of the techniques of transformations of the square root, logarithmic or reversing can be used. Sometimes it is necessary to apply several transformations to achieve normality of data.[4]

Of course, it should be noted that despite the more appropriate appearance of transformed data, it is more difficult to interpret the results and particular attention and expertise are required.[4] It is advisable to carry out all the analyses with the use of transformed data. However, finally, all the statistics should be converted to their initial form (the reverse of the transformation steps) to report the results.

Applying simple data cleaning and its effects: An oral health example

Finally, to review and practice the subjects discussed in this article on cleansing of data, the following data from a hypothetical research to evaluate the relationship between age and the decayed-missing-filled (DMF) index in a group of children under 15 years of age is evaluated here (Table 1).

To demine the relationship between the two variables of age and DMF in the samples mentioned above, the following path is used in SPSS:

Analyze → Correlate → Bivariate → Pearson

The calculated Pearson's correlation coefficient was -0.064 and non-significant (P = 0.788). The very small numeric value of 0.064 in relation to the correlation coefficient indicates a lack of relationship between age and DMF in the data available.

Based on what was discussed in this study, the prerequisite for the analysis of data is making sure that the data are clean. To this end, we return to the algorithm in the first section of the article and carry out the cleansing steps of data in the table 1.

Table 1. Row data from a hypothetical research to evaluate the relationship between age and the decayed-missing-fi (DMF) index

ID	Age (year)	DMF index
1	7	1
2	7	0
3	10	4
4	12	8
5	8	4
6	8	3
7	9	8
8	8	4
9	13	9
10	60	4
11	12	8
12	12	9
13	11	9
14	15	10
15	10	33
16	9	18
17	8	5
18	13	7
19	12	7
20	11	6

The first step is to evaluate data in relation to the presence of inconsistent data. In the column of age, the value 60 is seen! Since the study was carried out on children, the data on the age of the subject with an ID of 10 (ID = 10) has been entered incorrectly. Imagine that by re-evaluation of data, it becomes clear that the correct age of this individual is 10. Do you see any other inconsistent data in this table? By accurately evaluating the DMF column, it becomes evident that one of the data has been recorded as 33. Since the maximum of DMF in each individual is 28 or 32 (its maximum equals the number of teeth, i.e., 28, without taking wisdom teeth into account). Therefore, the value 33 cannot be correct. Imagine that by re-evaluation of the documents it becomes clear that the correct value is 7 and the correction is made.

The second stage of cleansing the data based on the algorithm is to evaluate outlier data, which can be accomplished through the following steps in SPSS:

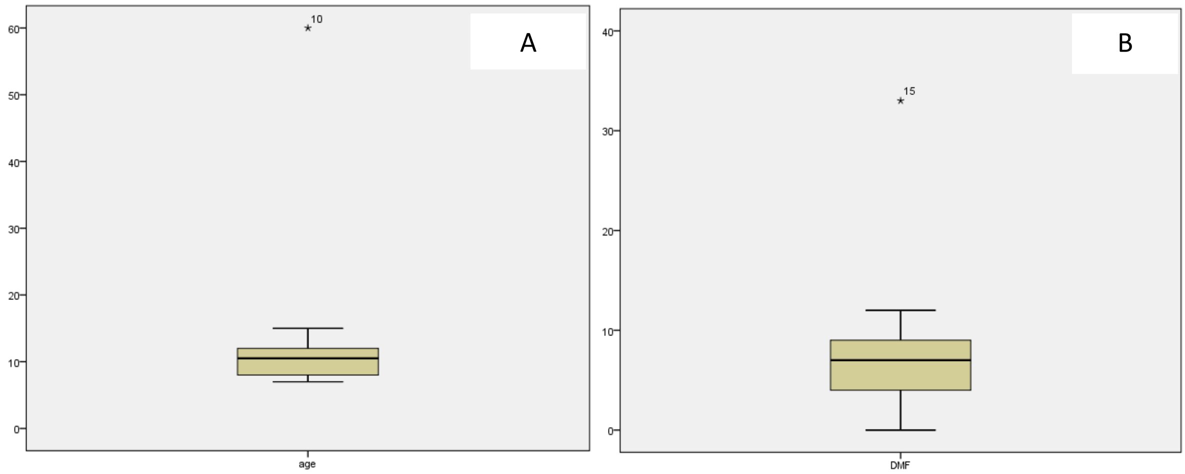

Figure 7. Outlier data showed in box plots [A for age and B for decayed-missing-filled (DMF)]

Analyze → Descriptive Statistics → Explore → Statistics → Outliers

In the following diagrams, outlier data are seen in the data on age and DMF (Figure 7):

As it is evident, in the data on age, the data of the ID = 10 is considered outlier, which was 60 as seen in the preceding paragraph and had been entered erroneously and was corrected. In the data on DMF, the data of ID = 15 is distant from other data. The DMF of ID = 15 is considered outlier and the value that had been erroneously entered was 33, which was corrected. The next stage in cleansing data is to evaluate missing data. In the data presented, in table 1, no data is missing. Of course, this is natural in this data series because missing data usually occur in studies with a very large sample size or in cases in which sensitive questions are asked during the course of the study or in situations in which the subjects are followed over a long period of time and the subjects may not show up in follow-up sessions; the data set here has none of the properties mentioned above.

After the data mentioned above were corrected due to inconsistency, the correlation coefficient of the two variables was calculated at 0.692, which is significant (P = 0.001). Note that changing only two variables in the table can significantly affect the results of the study.

Conclusion

Cleansing data before statistical analyses is necessary before drawing any conclusions from data. Researchers should make sure that no mistakes have been made in entering data in datasheets and after it in entering data into the statistical software. Evaluation of the presence of outlier data of one or several variables, missing data and normality of data distribution and if necessary, the transformation of data, are the principal steps in cleansing and preparation of data for analysis. It is suggested that the cleansing and preparation steps of data be explained in the final report of research studies.

Conflict of Interests

Authors have no conflict of interest.

Acknowledgments

We would like to show our gratitude to the Dr. Mohammad Reza Baneshi and Mrs. Maryam Hadipour for their valuable comments.

References

1. Szklo M, Nieto J. Epidemiology: beyond the basics. 3rd ed. Burlington, MA: Jones and Bartlett Learning; 2014.

2. Barchard KA, Pace LA. Preventing human error: The impact of data entry methods on data accuracy and statistical results. Comput Human Behav 2011; 27(5): 1834-9.

3. Van den Broeck J, Cunningham SA, Eeckels R, Herbst K. Data cleaning: detecting, diagnosing, and editing data abnormalities. PLoS Med 2005; 2(10): e267.

4. Peat J, Barton B. Medical statistics: a guide to data analysis and critical appraisal. 1st ed. Hoboken, NJ: Wiley Blackwell/BMJ Books; 2005. p. 338.

5. Barnett V, Lewis T. Outliers in Statistical Dat. New York, NY: Willey; 1994.

6. Osborne JW. Data cleaning basics: best practices in dealing with extreme scores. Newborn Infant Nurs Rev 2010; 10(1): 37-43.

7. Osborne JW, Overbay A. The power of outliers (and why researchers should always check for them). Pract Assess Res Eval 2004; 9(6): 1–12.

8. Hawkins D. Identification of outliers. New York, NY: Springer; 1980.

9. Selst MV, Jolicoeur P. A solution to the effect of sample size on outlier elimination. Q J Exp Psychol A 1994; 47(3): 631-50.

10. Iglewicz B, Hoaglin DC. How to detect and handle outliers. Milwaukee, WI: ASQC Quality Press; 1993.

11. Babaee G, Amani F, Biglarian A, Keshavarz M. Detection of outliers methods in medical studies. Tehran Univ Med J 2007; 65(7): 24-7. [In Persian].

12. Hamilton LC. Regression with Graphics: A Second Course In Applied Statistics. 1st ed. Belmont, CA: Duxbury Press; 1991.

13. Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ 2009; 338: b2393.

14. Baneshi MR, Talei AR. Impact of imputation of missing data on estimation of survival rates: an example in breast cancer. Iran J Cancer Prev 2010; 3(3): 127-31.

15. Pigott TD. A review of methods for missing data. Educ Res Eval 2001; 7(4): 353-83.

16. Ibrahim JG, Chen MH, Lipsitz SR, Herring AH. Missing-data methods for generalized linear models. J Am Stat Assoc 2005; 100(469): 332-46.

17. Donders AR, van der Heijden GJ, Stijnen T, Moons KG. Review: a gentle introduction to imputation of missing values. J Clin Epidemiol 2006; 59(10): 1087-91.

18. Tabachnick BG, Fidell LS. Using multivariate statistics. 6th ed. Boston, MA: Boston, Allyn and Bacon; 2012. p. 1024.

19. Dong Y, Peng CY. Principled missing data methods for researchers. Springerplus 2013; 2(1): 222.

20. Park HM. Univariate analysis and normality test using SAS, Stata, and SPSS. Technical Working Paper. Bloomington, IN: The University Information TechnologyServices (UITS) Center for Statistical and Mathematical Computing, Indiana University; 2008.

21. Doornik JA, Hansen H. An omnibus test for univariate and multivariate normality. Oxf Bull Econ Stat 2008; 70(1): 927-39.

22. Bulmer MG. Principles of Statistics. New York, NY: Dover Publications; 1979.