Fundamentals in Biostatistics for Investigation in Pediatric Dentistry: Part II –Biostatistical Methods

Amaury Pozos-Guillén[*] / Socorro Ruiz-Rodríguez[**]/ Arturo Garrocho-Rangel [***]

The main purpose of the second part of this series was to provide the reader with some basic aspects of the most common biostatistical methods employed in health sciences, in order to better understand the validity, significance and reliability of the results from any article on Pediatric Dentistry. Currently, as mentioned in the first paper, Pediatric Dentists need basic biostatistical knowledge to be able to apply it when critically appraise a dental article during the Evidence-based Dentistry (EBD) process, or when participating in the development of a clinical study with dental pediatric patients. The EBD process provides a systematic approach of collecting, review and analyze current and relevant published evidence about oral health care in order to answer a particular clinical question; then this evidence should be applied in everyday practice.

This second report describes the most commonly used statistical methods for analyzing and interpret collected data, and the methodological criteria to be considered when choosing the most appropriate tests for a specific study. These are available to Pediatric Dentistry practicants interested in reading or designing original clinical or epidemiological studies.

Keywords: *Evidence-based Dentistry, Biostatistics, Pediatric Dentistry, Dental Research.*

INTRODUCTION

According to John C. Pezzullo, in his book *Biostatistics for Dummies*, Biostatistics "*deals with the design and execution of scientific experimentson living creatures (human beings or animals), the acquisition and analysis of data from those experiments, and the interpretation and presentation of the results of those analyses*".[1] Statistical methods are essential tools for drawing valid conclusions through special techniques dealing with numerical data, that are variable among individuals.[2] Biostatistics methods consist predominantly of diverse steps, such as the generation of hypotheses, collection, analysis, and interpretation of the results in clinical terms, in order to accept or reject those hypotheses, with a high level of probability or security.[3,4]

Jacobson and Rowland, on the other hand, appropriately have noted that "*some readers gloss over the statistics found in research articles, trusting that the authors have 'done it right', but this is a lost opportunity to better understand the clinical significance and reliability of an article's conclusions*".[5] And this is finally the most important purpose of practicing the EBD philosophy.

In this context, the present paper claims to describe as clear as possible those common and practical statistical aspects, in order to understand the most frequent methods used for analyzing and interpreting numerical data, collected from clinical or epidemiological studies in Pediatric Dentistry. So, the reader will be able to better judge the validity of any paper and not simply takes for granted that their clinical results or findings were properly obtained by the authors.

Common statistical tests

Among the extensive statistical tests reported in the dental literature, we will describe only those that are most frequently employed, according to our experience as biostatistical paper/book readers and Pediatric Dentistry researchers. Selection of an adequate test primarily depends on the research question, the posed hypothesis,

From the Faculty of Dentistry, San Luis Potosi University, San Luis Potosí, SLP, México.

[*] Amaury Pozos-Guillén, DDS, PhD, Associate Professor, Basic Sciences Laboratory.

[**] Socorro Ruiz-Rodríguez, DDS, MS, Associate Professor, Pediatric Dentistry Postgraduate Program.

[***] Arturo Garrocho-Rangel, DDS, PhD, Associate Professor, Pediatric Dentistry Postgraduate Program.

Send all correpondence to:
Amaury Pozos Guillén
Facultad de Estomatología, Universidad Autónoma de San Luis Potosí. Av. Dr. Manuel Nava #2, Zona Universitaria, CP 78290, San Luis Potosí, SLP, México.
Phone: 52 444 8262300 X 5134
E-mail: apozos@uaslp.mx

and the type of data collected along the investigation. Usually, investigators are interested in discerning the difference between two or more means or proportions (comparative studies), or the associations and/or correlations between two or more variables.[6,7]

Choice of the statistical method

Statistical methods are essential to produce valid conclusions from the analyzed data. In general, the selection process of a statistical test, must be performed during the planning of the study; it mainly depends on the nature of the investigation question and design, and also the type and characteristics of the data that have been collected. As a broad guide (Fig.1), the following five questions should be answered when choosing the proper statistical test.[8,9]

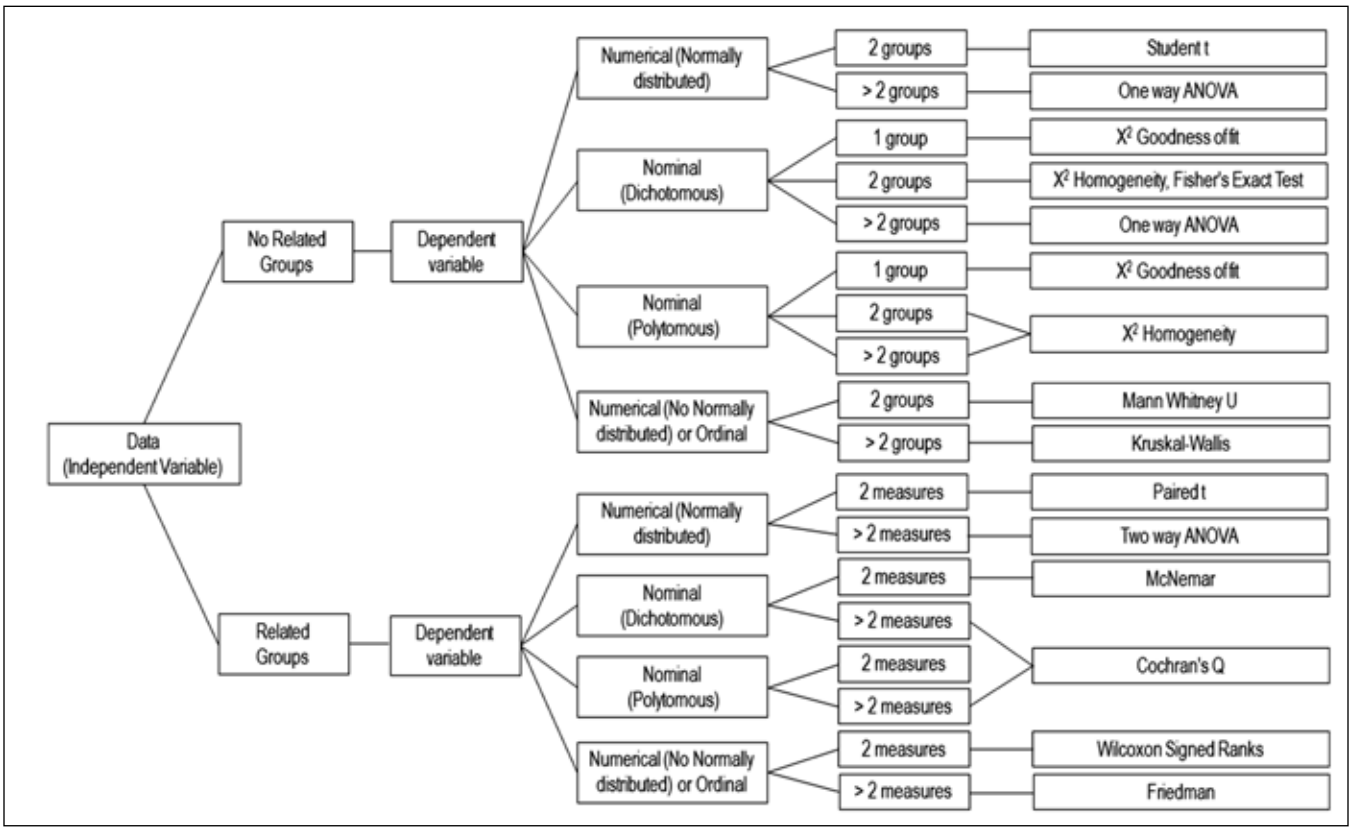
1. Is the hypothesis about an association or correlation (relationship) between variables or a difference (comparison) between groups?
2. Which is the data type (variable), categorical (nominal or ordinal) or quantitative?
3. Are the parametric test assumptions met (e. g. normally distributed)? Remember the parametric and non-parametric tests (see also below).
4. Are the studied groups or data related or dependent each other, or not? The term 'dependent' means that the subjects from *a single* sample were measured at *baseline* and *after* treatment, or pre- and post- dental procedures.[10] When two sample subjects are studied, they are 'independent' groups.
5. How many groups are being studied, two or more than two?; for example, a researcher can compare a new experimental procedure (group 1) vs. the best recognized procedure, as control (group 2) vs. a placebo (group 3). When testing more than two groups, the significance level must be adjusted (see below).

Regarding point 3, one manner of assuring 'normality' for data in an empirical way is by taking sample sizes with a number of > 30 or 40. This is a statistical principle denominated *the central limit theorem*;[5] this statement says that a subject sample, or n, will be considered normal when is large enough, even though the original population from which this sample was selected is not normally distributed.[11] In clinical dentistry and health sciences practice, many populations exhibit a non-normal distribution; however, the larger the extracted sample size from them, the better distributed the sample is. Thus, the central limit theorem enables us to perform more powerful parametric tests.[10]

Parametric and non-parametric statistical tests

Parametric procedures assumed that the chosen sample of experimental subjects was drawn from a normal population; as seen before, with large enough sample sizes > 30 or 40, parametric tests can be used even when the data are not normally distributed.[11] These statistical tests are more powerful because they take all possible information of the sample, so they are more sensitive to detect a real difference for rejecting a null hypothesis, with less probability of committing error.

Figure 1. Algoritm for choosing the proper statistical test.



On the other hand, non-parametric methods are indicated for analyzing nominal or ordinal data from sample sizes, employing easier computations, and are less sensitive to measurement discrepancies regarding the parametric tests. In general, parametric procedures are considered more reliable for analyzing data, although the non-parametric ones are quite popular when are adequately indicated.[10]

Although there are several statistical tests for normality, which calculate the probability that a sample was drawn from a normal population, two are frequently used. The most recommended is the *Shapiro-Wilk* procedure, which works best for data sets < 50, but can be used in larger samples; and the less powerful *Kolmogorov-Smirnov* test, indicated when data sets are > 50.[11]

Next we present the most popular statistical methods used in Pediatric Dentistry and other Health Sciences to analyze various types of data.

Comparison of means between two groups, or among three or more groups

The Student t test

In order to compare two groups with normally distributed numerical data on a specific outcome variable, investigators must establish whether groups are or are not related (dependent data). In cases of non-related data – taken from different individuals –, we can compare the sample means of the two groups using the *Student unpaired t* test, the only acceptable parametric statistical test for this purpose.[12] With respect to related data – taken from the same subject (e.g., from one side and from the other side of the mouth, or on two occasions: before and after treatment) belonging to the sample–, the *paired* t *test* is indicated; this design is powerful because subjects serve as their own controls, thus eliminating individual variation and, in cases of biological measurements, individual variation is usually large. The paired *t* test only analyzes the mean of the differences between the paired measurements for each subject.[6,12]

We can also calculate the 95% CI and the *p* value for the difference between means in order to prove whether this difference is statistically significant. Remember: If the 95% CI on the difference value between two means, for normally distributed data, does not contain the zero value, then we conclude that this difference is significant, and the null hypothesis can be rejected.[6,7]

When the data of both groups are not considered normally distributed, then non-parametric tests, which compare medians instead of means, can be used. The *Mann-Whitney* U is an alternative test, and one that is nearly as sensitive, as the unpaired *t* test; for two groups with small numbers, this test is considered statistically more sensitive. In a paired two-group design, you should use the *Wilcoxon Signed Rank Test*, the non-parametric equivalent of the paired *t* test.[8,9]

One-tailed or two-tailed tests

When we compare two groups (A and B) measured by continuous data, the null hypothesis dictates that there is no significant difference between group means. This difference can be in either direction: (A > B or B > A); if an investigator is not absolutely sure of the direction of the difference or considers both possibilities, which is the usual case, the test of significance is named a *two-tailed* test ('A ≠ B'). But, if the same investigator wants to know in advance whether the difference is only in one direction, or if the investigator considers only one possibility ('higher than...' or 'lower than ...'), based on valid evidence or previous knowledge, then the test is called the *one-tailed* test. For example, a Pediatric Dentist wants to test a new pharmacological regimen for treating uncooperative young children. He has read in a recent paper that the average onset time to achieve an adequate sedation of a proven drug combination (called 'XY'), is 50 minutes in children < 36 month-old. Then, his alternative hypothesis is that the new sedative regimen will have an average onset time *shorter than* XY combination (50 minutes), and must employ a one-tailed test to contrast the hypothesis. In cases of hypothesis testing, the z value, a parameter necessary for the sample size calculation, is different, according to the type of test chosen (one or two-tailed), and the α significance level selected:

Values of Z		α level	
		0.01	0.05
Type of test	One-tailed	2.33	1.645
	Two-tailed	2.57	1.96

As you can see, the lower the α level, the higher the z value; this means that a larger sample size is required to detect higher significant differences.The majority of dental journals currently request the reporting of the exact two-tailed *p* value as a standard norm for hypothesis testing (avoid terms like *NS* or *non-significant*).[2,13]

The ANOVA test

In cases of comparing, at one time, three or more mean outcomes from normally distributed, unrelated groups, we employ a statistical test termed *Analysis of Variance* (ANOVA). Prior to carrying out an ANOVA test, we must draw assumptions that diverse groups were randomly defined and that they have similar levels of variation (*homogeneity of variances*). There are two types of ANOVA. The one-way ANOVA, when data are categorized in only one way or factor – gender, levels of drug dosage, or age groups; for example, when a group of researchers want to evaluate and compare the mean bonding strength of four different orthodontic bracket resins, by measuring the average maximum dislodge force in kg in each group.[10] The two-way ANOVA is employed when two combined categories or factors are implied in a response; for example, the same group of researchers wants to compare, in addition to the mean bonding strength among the four groups (factor A), the means between girls and boys (factor B) in each of the four groups.[6,7,9]

There is another application of ANOVA when the trial employs a sample in which multiple or repeated measurements are taken from each subject at predetermined intervals, that is, before, during and after receiving an intervention, in order to assess changes in a particular outcome over time; in this case, the analysis is termed *repeated measures* ANOVA.[9] This design is especially useful when, for example, a clinician wants to compare the therapeutic effect of a new analgesic drug with regard to another and administered to their pediatric patients after undergoing a primary tooth extraction, during the next 3 days. As with paired *t* test, there is dependency among the measurements, and error due to variations within each individual is dismissed, increasing the chances of observing real-effect differences between interventions – as, in this example, two analgesics.[12,13]

When we perform multiple mean comparison analysis through ANOVA, the test results only indicate whether there is a significant difference among all of the means (or at least between one pair of means), but it does not specifically establish between which pair of means this difference lies. Therefore, post-tests, denominated *multiple or pairwise comparison procedures*, are necessary to determine where the difference or differences between groups of comparison can lie. It is not to perform pairwise t tests to compare three or more means at a time, because the probability of obtaining a significant result only by chance is increased; for example, in the previous mentioned case about comparing the mean bonding strength among of four independent groups, all at $\alpha = 0.05$, the probability of one or more significant results is 4 X .05 = 0.20, or 20%.[6] Diverse *post-hoc* tests have been suggested, such as the Bonferroni and Tukey methods.[6,10] These are briefly explained:

1. Bonferroni's method: Basically, it adjusts the original α level, dividing α by the number of comparisons made; in this case, 0.05/4 = 0.0125, so each pair comparison must be significant at the .0125 level, in order to be declared statistically different.
2. Tukey's method: Also named *honestly significant difference* (*HSD*) test. It is applicable when the sample sizes for the groups are equal. Although easy, the whole procedure can be reviewed at Dawson *et al.* (2nd ed, 2004, pages 134-5).[6]

Before applying ANOVA and multiple comparison tests, it is advised to carry out prior procedures to assure their indications; for example, visual inspection of the distribution or statistical tests for normality, and the *Levene* test for homogeneity of variances, to have a clear idea about complying with the assumptions previously mentioned. When these assumptions are not met, we may attempt to transform the numerical data; sometimes, simple logarithmic scale transformation is sufficient to obtain normality and equal variances.[6,7]

Alternatively, we may employ non-parametric methods, although they are less powerful.[10] The non-parametric alternative test for one-way ANOVA is the *Kruskal-Wallis test* for more than two unpaired group medians, in which the outcome variable is ranked in a few categories. In the cases of three or more related samples, as in repeated measures, the non-parametric procedure indicated is the *Friedman two-way ANOVA* test, for comparing medians. Likewise, there are multiple comparison (*post hoc* or comparison between pairs) methods for non-parametric tests, such as follow-up procedures to make pair-wise comparisons; most common *post hoc* techniques used in dentistry are the Mann-Whitney procedure (for independent samples), the Wilcoxon test (for paired samples or repeated measures).[6,7,9]

Comparing two or three or more group proportions

The Chi-square test

This versatile test is the most common method for analyzing proportions from two or more groups. The test is not based on any assumption regarding a distribution of any variable and is considered as a non-parametric method. *Chi-square* tests are appropriate for answering two types of study questions.[6,7,12]

- Is there a significant difference between the proportions measured from a multinomial and categorical variable in a single sample? Here, we use the test called *goodness of fit*. The null hypothesis is: "None of the proportions is equal to the others".
- Is there a significant dependence or association between the proportion of subjects with (or without)… vs. the proportion of those with (or without)…? In this case, we can use the chi-square to test *independence* or *association*, without considering a cause-effect relationship between two categorical variables involving two or more samples. Here, the null hypothesis to test is the following: "There is no relationship between the two variables, or the sample proportions are equal".
- Is there a significant risk-effect association between the presence or absence of this variable… and the occurrence or non-occurrence of this one…? Here, we want to assess whether the outcome is or is not dependent on, or influenced by, a causal or risk factor, or whether the outcome is or is not equally influenced by two different treatments; thus, we should employ a chi-square of *homogeneity*. This test generally utilizes dichotomous data, where there are only two samples, each divided into two classes, arranged in four cells or a *2 × 2 contingency table*. In cases of 2 × 2 contingency tables in which sample size is a number < 20–40, a more accurate alternative method called the *Fisher exact test* should be performed. In this case, the null hypothesis is expressed as "the risks are the same in the two samples" or that the two samples are homogeneous with respect to the risk factor-of-interest.

The Chi-square (X^2) test is applied for analyzing two or more dichotomous or polychotomous nominal variables expressed as the frequency or proportion of subjects belonging to any of the response categories. In cases in which subjects are paired, or in before-after treatment studies, and when the characteristic is nominal, the *McNemar test* for paired proportions is indicated; if there are more than two measurements in the same subject, we employ the *Cochran Q test.*[6,9]

Correlation and regression

As we stated in the first part of this series,[10] correlation is defined as interdependence, or the degree to which a quantitative variable increases or decreases as the other quantitative variable also changes or, in other words, the examined variables 'go together', where one variable is considered independent (denoted as 'X') and the other is dependent ('Y'). The main goal of a correlation analysis is to establish the direction and strength of the association between X and Y, both measured independently in the same study subject.[10] A positive direction means that as X increases, so does Y, or that both decrease together; for example, if the daily sugar intake in mg (X) increases, then the number of decayed teeth (Y) also increases. A negative direction occurs when X increases and Y decreases (or vice versa), as in the following example: if sugar intake increases, the number of healthy teeth decreases. The data point distribution may be shown in a graph called a *scatterplot.*[6,12]

The strength of a correlation is measured by means of a statistic known as the *Pearson correlation coefficient* (symbolized by '*r*'), in which values range from –1 to +1, or from a perfect negative to a perfect positive correlation and, to the degree that zero is closer to *r*, the correlation degree or dependence is smaller. The square of *r*, or 'r^2', is known as the *coefficient of determination*, which is a

statistical measure that 'explains' how much the variability of Y is explained by its relationship with X, as a value between 0 and 100%; the higher the value of r^2, the more the variation in Y is explained by X, and the better the degree of correlation between the variables. For example, if r^2 is 74.6%, it means that 74.6% of the variation in the dependent variable is caused by the independent variable, and the remainder (25.4%) is explained by other factors. In practice, values of r^2 that are >10% can be clinically significant.[6,7]

The simplest correlation technique involves only two variables – the *simple* correlation. The Pearson correlation is used when both variables are normally distributed; if the distribution is skewed by extreme values, then we apply a non-parametric alternative called the *Spearman test*. However, it is possible to analyze two or more independent variables vs. a dependent one –the *multiple* correlation- should be used.[7,12,14] It is noteworthy that correlation does not necessarily demonstrate that one variable causes the other; a statistical correlation does not imply causation.[14–16]

While the correlation analysis demonstrates the association magnitude between two variables, the *regression* analysis is able to predict the value of the dependent variable from knowledge of the value of the independent variable; this method is also named *simple linear regression.*[17] The term 'linear' refers to the fact that the relationship between X and Y is linear, in the manner of a straight line –the *regression line*–that best describes the relationship or correlation. The regression line is drawn through the points representing the values of both variables, minimizing the distance between the line and all of the data points (the *best fit* to the points), and indicates the direction of the correlation (positive or negative).[17] During regression analysis, we develop, by means of elementary geometry, the mathematical equation for any straight line that best describes the variable relationship: Y = a + bx, where 'a' is the *intercept*, the point at which the straight line crosses the Y axis, and 'b' is the *slope* of the line (or *regression coefficient*), which expresses the amount of increase (or decrease) of the Y value for each one-unit change in the X value. The sign, positive or negative, of b indicates the direction of the correlation.[6,10,14–16]

Types of statistical errors in dental clinical trials

In case of comparative studies – such as clinical trials – for determining differences between treatment effect means or proportions, investigators must consider the possibility of reporting false conclusions from their studies, mainly because of two types of error.[6,7,18]

- Type I error or α occurs in cases in which authors conclude that a significant difference exists when it does not – a false positive result.
- Type II error or β concludes that no difference exists when a real difference is present – a false negative result. The probability of finding such a real difference is called statistical power (or 1–β).

The best method for decreasing the possibility of committing these errors prior to data gathering is through careful sample-size calculation, which often takes into account in advance both the level of significance and statistical power values; values most commonly recommended are 0.05 and 0.20, respectively.[6,7,18–20]

Useful software for statistical analysis in Dentistry

Several suitable statistical programs employed for performing the most common tests are currently available. Some of the most commonly used for data analysis, data management and graphics, are: SAS (Statistical Analysis System), SPSS (Statistical Package for Social Sciences, by IBM®), R, Epi Info, and STATA; some of them are free of charge.[8,21]

CONCLUSIONS

The use of statistical methods for data analysis and interpretation are increasingly becoming an essential part in the process of providing dental care to children and adolescents, based on the philosophy of EBD, both in public and private health practices throughout the world. However, it has estimated that approximately half of the medical or dental papers contain statistical errors, some of which are serious enough to result in misleading conclusions. So, those Pediatric Dentists who desire to make and apply proper evidence-based clinical decisions to manage their patients should be able to critically judge, interpret and implement valid findings from the latest dental investigation, provided these findings are supported, among other issues, by a suitable statistical analysis.

REFERENCES

1. Pezullo JC. Biostatistics for dummies. 1st ed, pp 31–38. Hoboken: John Wiley & Sons, Inc., 2013.
2. Dakhale GN, Hiware SK, Shinde AT, Mahatme MS. Basic Biostatistics for post-graduate students. Indian J Pharmacol; 44: 435–442, 2012.
3. Mandrekar JN, Mandreaker SJ. Biostatistics: a toolkit for exploration, validation and interpretation of clinical data. J Thorac Oncol; 4: 1447–1449, 2009.
4. Hannigan A, Lynch CD. Statistical methodology in oral and dental research: pitfalls and recommendations. J Dent; 41: 385–392, 2013.
5. Jacobson BC, Rowland DY. A simpler approach to Biostatistics.GastrointestEndosc; 73: 570–574, 2011.
6. Dawson B, Trapp R. Basic and Clinical Biostatistics. 2nd ed, pp 82–97. New York: McGraw-Hill, 2004.
7. D'Agostino RB Sr, Sullivan LM, Beiser AS. Introductory Applied Biostatistics. 1st ed, pp 157–162. Belmont: Thomson Brooks/Cole, 2005.
8. Kyrgidis A, Triardis A. Methods and Biostatistics: a concise guide for peer reviewers. Hippokratia; 14(Suppl 1): 13–22, 2010.
9. Krzywinski M, Altman N. Nonparametric tests. Nat Methods; 11: 467–468, 2014.
10. Kim JS, Railey RJ. Biostatistics for Oral Healthcare.Ed. Blackwell Munksgaard, Oxford, UK; 1, 2008.
11. Ghasemi A, Zahediasl S. Normality tests for statistical analysis: A guide for non-staticians. Int J Endocrinol Metab; 10: 486–489, 2012.
12. Krzywinski M, Altman N. Significance, p values and t-tests. Nat Methods; 10: 1041–1042, 2013.
13. Krzywinski M, Altman N. Power and sample size. Nat Methods; 10: 1139–1040, 2013.
14. Krithikadatta J, Valarmathi S. Research methodology in dentistry: Part II – The relevance of statistics in research. J Conserv Dent; 15: 206–213, 2012.
15. Petrie A, Bulman JS, Osborn JF. Further statistics in dentistry. Part 1: Research designs 1. Br Dent J; 193: 377–380, 2002.
16. Williams AC, Bower EJ, Newton JT. Research in primary dental care. Part 6: Data analysis. Br Dent J; 197: 67–73, 2004.
17. Petrie A, Bulman JS, Osborn JF. Further statistics in dentistry. Part 6: Multiple linear regression. Br Dent J; 193: 675–82, 2002.
18. Sedgwick P. Pitfalls of statistical hypothesis testing: Type I and type II errors. BMJ; 349: g4287, 2014.
19. Morton JP. Reviewing scientific manuscripts: how much statistical knowledge should a reviewer really know? Adv Physiol Educ; 33: 7–9, 2009.
20. Pezullo JC. Biostatistics for dummies. 1st ed, pp 51–60. Hoboken: John Wiley & Sons, Inc., 2013.
21. Dembe AE, Partridge JS, Geist LC. Statistical software applications used in health services research: Analysis of published studies in the U. S. BMC Health Serv Res; 11: 252–257, 2011.