Med Pregl 2024; LXXVII (9-12): 325-330. Novi Sad: septembar-decembar.

PROFFESIONAL ARTICLES
STRUČNI ČLANCI

VARIABLES, DATA, AND DESCRIPTIVE STATISTICAL METHODS – PART I

VARIJABLE, PODACI I DESKRIPTIVNE STATISTIČKE METODE – DEO I

Božana NIKOLIĆ[1,2] and Tamara POPOVIĆ[3]

ORCID NUMBER
Božana Nikolić – 0000-0002-3468-6296
Tamara Popović – 0009-0001-0113-4727

University of Novi Sad, Faculty of Medicine Novi Sad, Department of Pharmacy, Novi Sad[1]
Health Center "Novi Sad", Novi Sad[2]
University Clinical Center of Vojvodina, Clinic for Psychiatry, Novi Sad[3]

Proffesional article
Stručni članak
UDK 001.8
https://doi.org/10.2298/MPNS2412325N

Abstract
Introduction. Statistics play a crucial role in clinical practice, enabling the comprehension of professional and scientific literature. A fundamental understanding of variables, data, and both descriptive and inferential statistics is essential. **Variables.** Variables represent characteristics that differ between subjects. They are measured, and their values are expressed by data. **Scales of measurement**. There are four measurement scales: nominal, ordinal, interval, and ratio. The nominal scale classifies qualitative variable into distinct categories without any inherent order. The ordinal scale categorizes qualitative variables in a specific sequence, though the differences between categories lack a defined numerical meaning. The interval scale applies to quantitative variables, ensuring that the differences between successive values are equal, but it lacks an absolute zero. The ratio scale also features equal intervals between values but includes an absolute zero, allowing for meaningful ratio comparisons. **Descriptive statistics**. Descriptive statistics summarize and numerically represent data to enhance clarity and interpretability. This includes measures of frequency, central tendency, variability, and distribution shape, alongside tabular and graphical data presentations. **Conclusion**. This article outlines fundamental concepts related to variables, data, and descriptive statistics, providing practical examples to illustrate their application.
Key words: Statistics as Topic; Data Collection; Normal Distribution; Clinical Medicine

Sažetak
Uvod. Statistika je neophodna u kliničkoj praksi zbog praćenja stručne i naučne literature. Stoga, treba poznavati osnovne koncepte povezane sa varijablama, podacima, deskriptivnom te inferencijalnom statistikom. **Varijable.** Varijable su obeležja koja se mogu menjati od jednog ispitanika do drugog. Mere se i njihove vrednosti se izražavaju podacima. **Skale merenja.** Postoje četiri skale merenja: nominalna, ordinalna, intervalna i skala odnosa. Nominalna skala samo rangira kvalitativnu varijablu u kategorije bez bilo kakvog redosleda. Ordinalna skala rangira kvalitativnu varijablu u kategorije po određenom redosledu, pri čemu rastojanja između kategorija nemaju precizno značenje. Intervalna skala podrazumeva da su rastojanja između uzastopnih vrednosti kvantitativne varijable uvek jednaka, ali da ne postoji apsolutna nulta tačka. Skala odnosa, pored jednakosti rastojanja između uzastopnih vrednosti kvantitativne varijable, ima i apsolutnu nulu. **Deskriptivna statistika.** Deskriptivna statistika je oblast statistike koja se bavi sumiranjem i numeričkim opisivanjem podataka kako bi bili jasni i razumljivi. Numeričko opisivanje podataka se očituje u merama učestalosti, merama centralne tendencije, merama varijabilnosti, merama oblika raspodele te tabelarnom i grafičkom predstavljanju podataka. **Zaključak.** Ovaj stručni rad, kroz primere, opisuje osnovne koncepte koji se odnose na varijable, podatke i deskriptivnu statistiku.
Ključne reči: statistika; prikupljanje podataka; normalna distribucija; klinička medicina

Introduction

Statistics are an essential component of clinical research [1, 2]. But why is this the case? Why it is statistical analysis indispensable? Clinicians work with biological systems that are inherently variable. For instance, when healthy individuals are exposed to a virus such as severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), their response can vary significantly – from no infection to asymptomatic, mild or moderate infection, to severe disease with potentially fatal outcomes [3, 4]. Consider a scenario where an individual receives a vaccine prior to viral exposure and does not develop an infection. Given the natural variability in immune responses, it would be difficult to conclude that the vaccine was protective based on a single case. To address this, clinical studies compare two groups: one receiving the vaccine and the other not. By monitoring infection rates in both groups and applying statistical analysis, researchers can determine whether observed differences are due to chance or a true effect of the intervention.

Beyond research, statistical literacy is crucial for clinicians to understand and critically evaluate sci-

✉ Corresponding author: Božana Nikolić, E-mail: bozana.nikolic@mf.uns.ac.rs

Abbreviations

SARS-CoV-2 – Severe Acute Respiratory Syndrome Coronavirus-2
COVID-19 – Corona Virus Disease of 2019

entific literature [1]. Medical knowledge is rapidly evolving, with new diagnostic techniques, therapies, and even emerging – such as COVID-19 – continually reshaping clinical practice. Relying solely on textbook knowledge, while disregarding the latest clinical evidence, is professionally unacceptable. A lack of statistical understanding may hinder clinicians from correctly interpreting study findings, potentially preventing patients from receiving the most effective and safest treatment. Therefore, familiarity with fundamental statistical concepts – including variables, data types, descriptive and inferential statistics is essential [5].

However, research indicates that many clinicians have a limited grasp of basic statistical principles [6-8]. For example, a study by Alomi et al. found that pharmacists scored an average of 64.6% on statistical knowledge test [6]. Similarly, Collins et al. highlighted a lack of statistical proficiency among surgeons and emphasized the importance of continuing education [7]. In another study, Jahan et al. reported that 73.6% of primary care physicians rated their statistical knowledge at or below 5 on a 10-point scale and expressed interested in improving it [8].

This article provides an overview of fundamental concepts related to variables, data, and descriptive statistics.

Variables

Variables are characteristics that vary between subjects, with their values represented as data. For example, sex is a variable, while is the classification of an individual as male or female constitutes data. Before conducting a clinical study, it is essential to consider the nature and type of variables, as they influence how data is described and analyzed.

The two primary types of variables are qualitative and quantitative, as illustrated in **Figure 1** [9, 10].

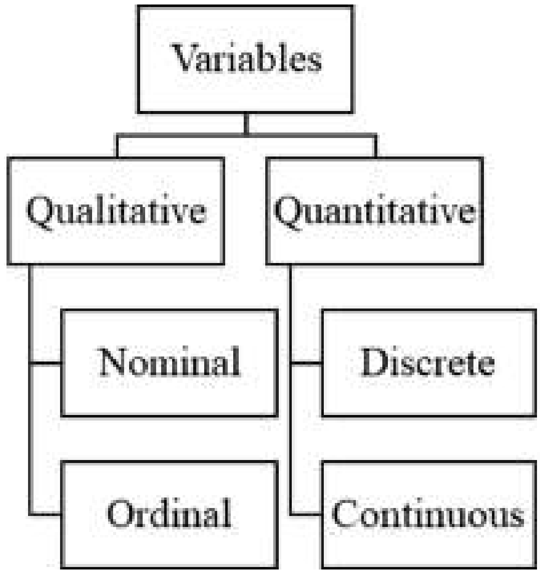

Figure 1. Types of variables

Qualitative variables categorize characteristics into distinct groups and are often referred to as categorical variables. Their measurement provides information about the frequency of characteristics within a group. These variables can be nominal or ordinal **(Figure 1).** Nominal variables classify data into categories without an inherent order, such as sex (male or female), marital status (single, married, or divorced), or presence of osteoporosis (yes or no). Ordinal variables have categories arranged in a meaningful order. Examples include educational level (elementary, high school, or college), pain intensity (rated on a 0-10 scale), and symptom frequency (never, rarely, sometimes, often, or always).

Quantitative variables are measured using a device or standardized evaluation method and are often termed numerical variables. They provide information about the quantity of a characteristic and can be classified as discrete or continuous **(Figure 1).** Discrete variables assume specific numerical values, typically integers. Examples include age, number of students in a classroom, and number of hospitalizations. Continuous variables can theoretically take any value within a given range. However, measurement precision is limited by the accuracy of the instrument. For instance, a patient's actual blood glucose level might be 5.2372 mmol/L, but a device may only record it as 5.2 mmol/L. Examples of continuous variables include body mass and height and blood glucose concentration.

In addition to their classification as qualitative or quantitative, variables are also categorized as dependent or independent based on their role in a study [10]. The dependent variable is the outcome being studied – it is a characteristic that researchers aim to explain or predict. The independent variable is the factor manipulated or observed to determine its effect on the dependent variable. For example, in a study examining the relationship between sex and cardiovascular morbidity, the dependent variable is cardiovascular morbidity, while the independent variable is sex. Typically, studies have a small number of dependent variables (one or, in some cases, a few), whereas the number of independent variables can be extensive.

Scales of measurement

There are four primary scales of measurement: nominal, ordinal, interval, and ratio [9, 10].

Nominal scale measurements classify qualitative characteristics into distinct categories without any inherent order. These categories are entirely different from one another, and any numerical or alphabetical coding assigned to them serves only as a label rather than representing actual quantities (e.g., 1 – female, 2 – male). A nominal scale with only two categories is referred to as a dichotomous or binary scale.

Ordinal scale measurements categorize qualitative characteristics in a ranked order, but the differences between categories do not have a precise numerical meaning. For example, pain intensity rated on a 0-10 scale follows an ordinal structure (category 2 < category 4 < category 8), but the difference between 4 and 8 does not imply that pain is twice as intense as category 4.

Interval scale measurements ensure that the difference between successive values is always equal, but they lack an absolute zero point, only an arbitrary one. A classic example is temperature measured in degrees Celsius, where the difference between 10°C and 20°C is the same as that between 30°C and 40°C; however, 40°C is not twice as hot as 20°C, as zero degrees does not represent the complete absence of heat.

Ratio scale measurements have an equal distance between successive values but also include an absolute zero, making meaningful ratio comparisons possible. For example, temperature in Kelvin is measured on a ratio scale, where 0 Kelvin represents the absence of molecular motion [10].

Categorical variables are measured using nominal or ordinal scales, whereas continuous variables are measured on an interval or ratio scales [10].

Descriptive statistics

Descriptive statistics summarize and numerically represent data, making statistical summarization essential for meaningful analysis.

Numerical data description is categorized into several groups: measures of frequency, measures of central tendency, measures of variability, measures of distribution shape, and tabular and graphical data presentation [9].

Measures of frequency

Measures of frequency indicate how often a value appears within a dataset. These include count, percentage and frequency, which are primarily used for categorical data measured on nominal and ordinal scales. For example, Nikolić et al. presented in Tables 2-5 in their article drug interactions data in terms of frequency and percentage [11].

Measures of central tendency

Measures of central tendency describe how data cluster around a central value, providing a representative metric for comparison across groups. The three most commonly used measures are the mean, median, and mode [12, 13].

The mean (arithmetic mean) is an appropriate measure for data measured on an interval or ratio scale with a normal distribution. It is calculated by dividing the sum of all values by the total number of observations. However, it is highly sensitive to outliers, i.e. data with extremely low or extremely high values.

The median is the middle value when data are arranged in ascending order. It is suitable for ordinal data, datasets with outliers, and continuous data that do not follow a normal distribution. It is the value that divides the series, where data is sorted by size, into two equal parts. There is an equal number of data (50%) above and below the median. If there is an odd number of observations, the median is the middle value. If there is an even number, it is the average of two central values.

The mode represents the most frequently occurring value in a dataset and is primarily used for nominal data. A variable may have one mode, two modes, multiple modes, or no mode at all.

Measures of variability

While measures of central tendency summarize data with a single representative value, they do not provide insight into how individual data points deviate from the mean or from each other. This variability, also referred to as dispersion, is crucial for understanding the distribution of data. Several statistical measures quantify variability, the most common being range, interquartile range, variance, and standard deviations [12, 13]. Broadly, range and interquartile range assess the spread of data points relative to each other, while variance and standard deviation measure how data points deviate from the mean.

Range is a measure of variability applicable to ordinal, interval, and ratio scale. It is calculated as the difference between the maximum and minimum values. However, due to its sensitivity to extreme values (outliers), the range is generally unsuitable for clinical studies, where outliers frequently occur.

Interquartile range is a measure of variability used for ordinal data and continuous data that do not follow a normal distribution. It always includes the median and represents the range within which the central 50% of data points lie. It is calculated as the dif-

Table 1. Graphical presentation of data: examples

Qualitative data	Quantitative data
Pie chart	Line diagram
Bar chart	Histogram
Pareto chart	Frequency polygon
	Distribution curve
	Box plot
	Scatter plot
	Survival curve

ference between the third quartile and the first quartile. The third quartile marks the 75th percentile, and the first quartile marks the 25th percentile. Because the interquartile range excludes extreme values, it is preferred over the range in many clinical applications. For instance, Zdravković et al. reported body weight, height, and body mass index as median and interquartile range in **Table 1** of their article [14].

The variance and standard deviation are key statistical measures used to describe data on an interval or ratio scale, particularly when the data follows a normal distribution. Variance represents the average squared deviation of individual values from the mean, while standard deviation, the most commonly used measure of variability, is mathematically derived as the square root of variance. In **Table 1** of the study by Krishnaveni et al., lipid profile data of patients are presented using mean and standard deviation [15].

While measures such as range, interquartile range, variance, and standard deviation express variability in the same units as the measured characteristic, they are not suitable for comparing the variability across characteristics with different units. In contrast, the coefficient of variation is a unit-free measure that allows such comparison (e.g., newborn body weight versus newborn body length). Defined as the ratio of standard deviation to the mean, it is typically expressed as a percentage. For instance, in **Table 1** of the study by Schmidt et al., glycemic control data are reported using the coefficient of variation [16]. Generally, a coefficient of variation of 30% or lower indicates statistical homogeneity, while values exceeding 30% suggest heterogeneity [17, 18].

Measures of distribution shape

Data distributions vary in shape, with normal distribution being the most significant for continuous variables [13]. Many biological processes and statistical analyses assume normality, graphically represented by a symmetric, bell-shaped curve in which the mean, median, and mode are identical **(Figure 2A)** [9].

Skewness and kurtosis are used to assess and describe the shape of distribution **(Figures 2 and 3)** [9].

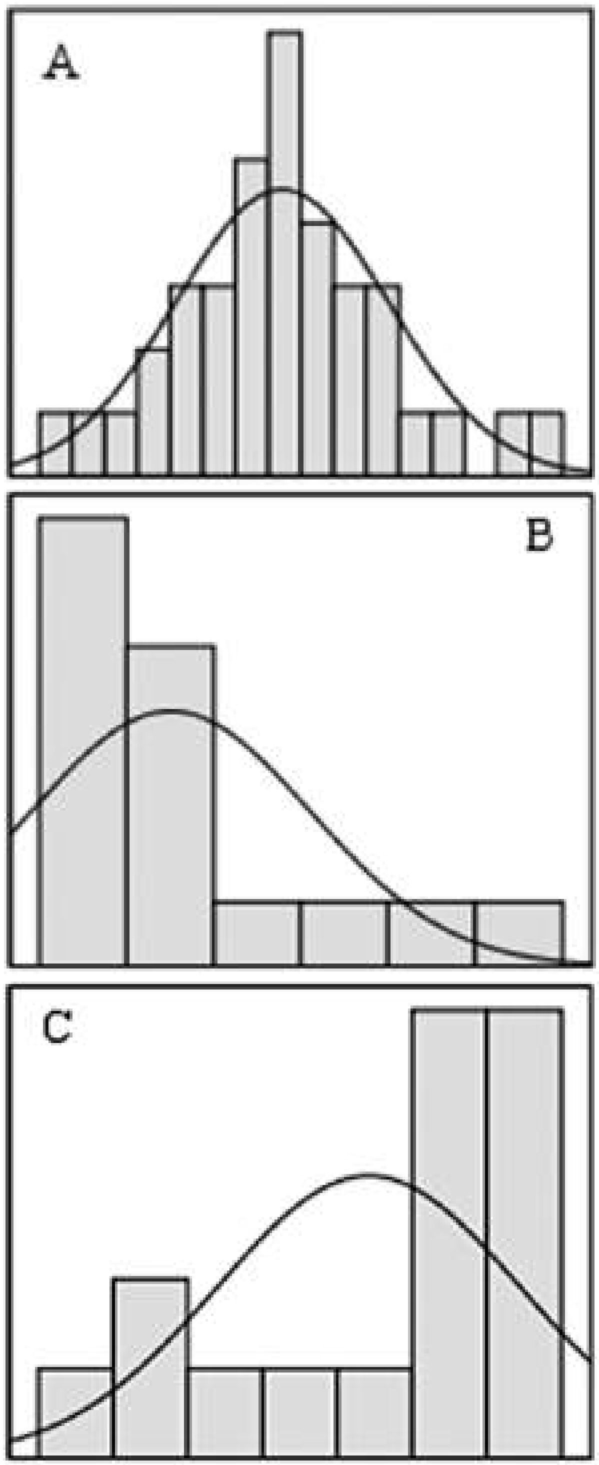

Figure 2. Distribution shape according to symmetry: A. normal (symmetric); B. positive skewed (asymmetric); C. negative skewed (asymmetric)

Skewness measures asymmetry in data distribution. A positively skewed distribution (right-skewed) occurs when the mode and median are less than the mean, shifting the peak leftward, with most values concentrated on the lower end and a longer right tail **(Figure 2B)**. Conversely, a negatively skewed distribu-

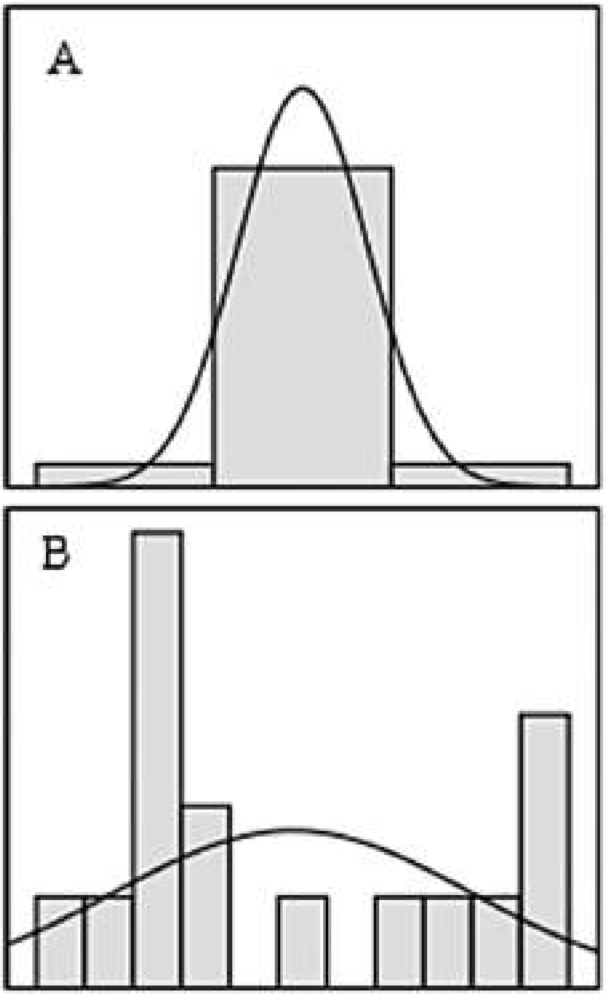

Figure 3. Distribution shape according to height and peak sharpness: A. positive kurtosis; B. negative kurtosis

tion (left-skewed) arises when the mode and median exceed the mean, shifting the peak rightward, with most values higher and a longer left tail **(Figure 2C).**

When skewness is approximately zero, the distribution is symmetric. In clinical studies, perfect symmetry is rare. A skewness range of -1 to +1 suggests minor deviation from normality, though values between -2 to +2 are often considered acceptable [17].

Kurtosis describes the height and sharpness of the distribution's peak. A positive kurtosis value indicates a high, narrow peak, while a negative value signifies a flatter, wider distribution **(Figure 3).**

As with skewness, kurtosis values between -1 and +1 suggest minor deviations from normality, though a range of -2 to +2 is generally acceptable [17].

Tabular and graphical data presentation

Once data have been collected and summarized, they should be presented clearly and concisely. This is typically achieved using tables and graphs. Tables provide the most comprehensive data representation and can be categorized by content (simple, complex, or combined) and purpose (reference and analytical). Graphs visually supplement tables rather than replace them. The choice of graphical representation depends on data type **(Table 1)** [9, 19]. Commonly used visualizations include histograms, bar charts, line charts, scatter plots, and box plots [20].

Conclusion

Statistics are fundamental to clinical research and essential for interpreting scientific literature. While an in-depth understanding of complex mathematical formulas is not required, researchers and clinicians must grasp key statistical concepts, including variable classification, data, scales of measurement, descriptive and inferential statistics.

References

1. Evans SR. Common statistical concerns in clinical trials. J Exp Stroke Transl Med. 2010;3(1):1-7.

2. Aggarwal R. Statistical literacy for healthcare professionals: why is it important? Ann Card Anaesth. 2018;21(4):349-50.

3. Pereira NL, Ahmad F, Byku M, Cummins NW, Morris AA, Owens A, et al. COVID-19: understanding inter-individual variability and implications for precision medicine. Mayo Clin Proc. 2021;96(2):446-63.

4. Yang KK, Paris CF, Gorman KT, Rattsev I, Yoo RH, Chen Y, et al. Factors associated with resistance to SARS-CoV-2 infection discovered using large-scale medical record data and machine learning. PLoS One. 2023;18(2):e0278466.

5. Popović J, Nikolić B, Popović K, Milošević N, Popović D. Statistika u kliničkim istraživanjima – interpretacija rezultata. In: Popović J, ed. Statistički principi u kliničkoj farmaciji. Novi Sad: Futura; 2021. p. 1-64.

6. Alomi YA, Altebainawi AF, Alharbi AA, Alabdullatif AA. Biostatistical analysis knowledge of pharmacy research in the Kingdom of Saudi Arabia. International Journal of Advanced and Applied Sciences. 2020;7(3):104-12.

7. Collins AP, McCall M, Cassinat J, Grise A, Schwartzman J, Kelly J, et al. Performance of orthopaedic shoulder and elbow surgeons on a biostatistical knowledge examination. Adv Orthop. 2023;2023:8840263.Jahan S, Al-Saigul AM, Suliman AA. Attitudes to statistics in primary health care physicians, Qassim province. Prim Health Care Res Dev. 2016;17(4):405-14.

8. Spasić S, Kotur-Stevuljević J, Bogavac-Stanojević N. Primenjena statistika za studente farmacije. Beograd: Univerzitet u Beogradu, Farmaceutski fakultet; 2020.

9. Walters RW. The application of statistical analysis in the biomedical sciences. In: Malone PM, Malone MJ, Witt BA, Peterson DM, editors. Drug information: a guide for pharmacists. 7th ed. New York: McGraw Hill; 2022. p. 289-396.

10. Nikolić B, Popović J, Bećarević M, Rakić D. Exposure to potential drug-antimicrobial agent interactions in primary health care. Vojnosanit Pregl. 2018;75(8):795-802.

11. De Sousa Rodrigues CF, de Lima FJC, Barbosa FT. Importance of using basic statistics adequately in clinical research. Rev Bras Anestesiol. 2017;67(6):619-25.

12. Ranganathan P, Gogtay NJ. An introduction to statistics – data types, distributions and summarizing data. Indian J Crit Care Med. 2019;23(Suppl 2):S169-70.

13. Zdravković R, Tatić M, Golubović M, Preveden M, Klinovski M, Vicković S. Effects of mechanical ventilation during cardiopulmonary bypass on postoperative pulmonary complications. Med Pregl. 2023;76(1-2):10-5.

14. Krishnaveni P, Gowda VM. Assessing the validity of Friedewald's Formula and Anandraja's formula for serum LDL-cholesterol calculation. J Clin Diagn Res. 2015;9(12):BC01-4.

Rad je primljen 4. I 2024.
Recenziran 24. II 2025.
Prihvaćen za štampu 24. II 2025.
BIBLID.0025-8105:(2024):LXXVII:9-12:325-330.

15. Schmidt S, Christensen MB, Serifovski N, Damm-Frydenberg C, Jensen JB, Fløyel T, et al. Low versus high carbohydrate diet in type 1 diabetes: a 12-week randomized open-label crossover study. Diabetes Obes Metab. 2019;21(7):1680-8.

16. Milošević Z, Bogdanović D. Statistika i informatika u oblasti medicinskih nauka. Niš: Galaksija; 2012.

17. Vaz MAB, Pacheco PS, Seidel EJ, Ansuj AP. Classification of the coefficient of variation to variables in beef cattle experiments. Ciencia Rural. 2017;47(11):e20160946.

18. Hazra A, Gogtay N. Biostatistics series module 1: basic of biostatistics. Indian J Dermatol. 2016;61(1):10-20.

19. Vetter TR. Descriptive statistics: reporting the answer to the 5 basic questions of who, what, why, when, wher, and a sixth, so what? Anesth Analg. 2017;125(5):1797-802.