JIRSS (2023)
Vol. 22, No. 01, pp 123-135
DOI: 10.22034/jirss.2023.707641

Statistical Relationship between Quantitative and Dichotomous Variables: Student's Test and Moving Average Approach

Anatoly Nikolaevich Varaksin [1], **Tatiana Anatolyevna Maslakova** [1], **Ekaterina Danilovna Konstantinova** [1], **Yulia Valeryevna Shalaumova** [2]

[1]Institute of Industrial Ecology, Ural Branch of the Russian Academy of Sciences, 620990, Ekaterinburg, Russia.
[2]Institute of Plant and Animal Ecology, Ural Branch of the Russian Academy of Sciences, 620144, Ekaterinburg, Russia.

Received: 28/03/2023, Accepted: 15/09/2023, Published online: 15/06/2024

Abstract. A new technique is proposed for evaluating the statistical relationship between a quantitative variable Y and a dichotomous variable X assuming two values: $X = 0$ and $X = 1$. The technique is based on the division of the quantitative variable Y into strata by the moving average technique and computation of average values in the strata for the variables Y and X. Stratification turns the dichotomous variable X into a quantitative one. Once the variable X has been transformed in this way, the statistical relationship between Y and X may be analyzed by linear regression and by analysis of variance. Thus, the technique proposed expands the range of methods available for analyzing statistical relationships between quantitative and dichotomous variables. Specific examples are used to compare the moving average technique with the t-test for symmetric (normal) and asymmetric distributions of quantitative variable Y. It is shown that the statistical relationship between stratified Y and X can be strongly different for a symmetrically (normally) distributed variable Y.

Keywords. Analysis of Variance, Body Mass Index, Linear Regression, Moving Average, Smoking, Stress Index, Student's Test, Quantitative Predictor and Dichotomous Outcome.

Anatoly Nikolaevich Varaksin (varaksin@ecko.uran.ru)
Tatiana Anatolyevna Maslakova (masta@ecko.uran.ru)
Ekaterina Danilovna Konstantinova (K_konst@ecko.uran.ru)
Corresponding Author: Yulia Valeryevna Shalaumova (yulyash@gmail.com)

MSC: 62D20, 62F03.

1 Introduction

There is a variety of statistical methods available for establishing the relationship between quantitative and dichotomous variables. For the purpose of this paper, these methods may be divided into two classes: class 1 - the quantitative variable is an outcome, and the dichotomous one is a predictor. In this class, use is made of the Student, Kolmogorov-Smirnov, Kramer-von Mises tests and others (Afifi et al. (1979), Büning (2002), Mosteller et al. (1977), Nuijten et al. (2016), Walker et al. (2010)). Class 2 - the quantitative variable is a predictor while the dichotomous one is an outcome. The best known method for establishing the relationship between such variables is logistic regression (Hosmer et al. (2013), Shoukri et al. (1999)). The paper considers the first class of methods where quantitative Y is an outcome and dichotomous X is a predictor.

Among the parametric methods for establishing statistical relationships between quantitative Y and dichotomous X, a special place belongs to Student's test, which is a special case of a general linear model (Mosteller et al. (1977), Student (1908), Zimmerman et al. (1992)). In particular, this method makes it possible to not only test the hypothesis of a statistical relationship between Y and X (i.e. to answer the question whether there are statistically significant differences between the average values of Y in two groups in relation to X) but also demonstrate the characteristics of the groups compared in a manner that is understandable to practicing clinicians, biologists, ecologists, that is to say in terms of group average values of Y and standard deviations (le Cessie et al. (2020)).

The t-test helps compare two average values of the quantitative variable Y in the groups (strata) defined by two values of the dichotomous variable: $\langle Y(X=0)\rangle$ and $\langle Y(X=1)\rangle$ (the parentheses stand for averaging). Statistically significant differences discovered between $\langle Y(X=0)\rangle$ and $\langle Y(X=1)\rangle$ point to a significant influence of X on Y (a relationship between Y and X).

Student's test has been used in numerous publications, whose authors emphasized the well-known fact: Student's test is applicable to quantitative variables Y where the Y values in groups $X = 0$ and $X = 1$ are normally distributed and sample size is sufficient (Hazra et al. (2016), Kim et al. (2019)). If there is a considerable departure from these assumptions, it is recommended using non-parametric tests (Afifi et al. (1979), Gerald et al., (2021), Hettmansperger (1984)). In this paper, the Student's method is compared with the moving average approach for both normally (symmetrically) and asymmetrically distributed outcomes Y.

2 The Moving Average Approach in the Case of Student Test

The moving average methods are widely used in statistical analysis, particularly in time series analysis for smoothing out random fluctuations in the factor studied and for highlighting trends. In time series, we have pairs of observations (ti, Yi), $i = 1, 2, \ldots n$, where the horizontal axis t is ordered time, and Y (the vertical axis) is a considerably scattered quantitative variable. Mosteller et al. (1977), Tukey (1977) suggested choosing the smoothing window size as $n_{\mathrm{w}} = 3$. In this case, at the first stage we select the first three values of time t $(t_1, t_2, and\ t_3)$, for which we calculate the average over three values of Y $(Y_1, Y_2, and\ Y_3)$ corresponding to the three first values of t (which provides the first stratum). At the next stage of the smoothing process, we select three values of t $(t_2, t_3\ and\ t_4)$ shifted with respect to the first stratum by one step of t. For these three values of t, we calculate the average of the three values of Y $(Y_2, Y_3\ and\ Y_4)$. The result is a second stratum with its own average values of t and Y. The smoothing operation is performed consistently as far as the last value of t. As a result, we obtain $(n - n_{\mathrm{w}} + 1)$ strata, in each of which we have average values for t and Y (n_{w} is the smoothing window, n is the number of observations).

Thus, as the moving average operation proceeds with respect to t and Y, the scatter of Y obviously decreases (the time series is smoothed). In the case of Student's test, the situation is different. We have a quantitative outcome Y and a dichotomous predictor X ($X = 0$ or $X = 1$), the pairs $(X_{\mathrm{i}}, Y_{\mathrm{i}})$ for $i = 1, 2, \ldots n$ being independent observations (not time series). To perform the moving average operation, we order the values of the outcome Y and then divide these Y values into strata in the same way as described above for time series. In each stratum, we calculate the average values of X and Y. Following this operation, the dichotomous predictor X is not smoothed (scatter of X from 0 to 1 remains the same). Rather, this dichotomous predictor acquires a range of new values in the interval from 0 to 1. Thus, if the smoothing window size is equal to , then the moving average operation provides a set containing $(n_{\mathrm{w}} + 1)$ values of X: $(0; 1/n_{\mathrm{w}}; 2/n_{\mathrm{w}}; \ldots..; (n_{\mathrm{w}} - 1)/n_{\mathrm{w}}; 1)$.

Our purpose in carrying out the moving average operation is to turn the dichotomous predictor X into a quantitative variable. It becomes then clear that the window size n_{w} cannot be too small because in this case the number of different values which the dichotomous variable X can assume is low and the smoothing (averaging) of the quantitative variable Y is weak. For instance, for $n_{\mathrm{w}} = 3$, the moving average procedure results in the dichotomous variable X assuming only four values: 0, 1/3; 2/3 and 1. If, however, the window size n_{w} is very large, the extent of averaging over the quantitative variable Y would be excessive.

In our studies, we use a window size of $n_{\mathrm{w}} = 10$, for which the transformed variable X assumes values from zero to one at a step of 0.1 (11 possible values of X), or $n_{\mathrm{w}} = 20$, for which the variable X takes values from zero to one at a step of 0.05 (21 possible values of X). If n_{w} is sufficiently large (say, $n_{\mathrm{w}} \geqslant 10$), the dichotomous predictor X turns (actually) into a quantitative variable. As the result of applying the moving average, we obtain two quantitative variables (averaged values of Y and X), the relations between which could be analyzed with the help of linear regression methods.

The moving average technique is employed in various software, such as SAS and SPSS (Krispin (2019), Yaffee et al. (2000), Yaffee (2007)). In this paper, we computed the moving averages in Statistica for Windows using the computer program (Varaksin et al. (2022)).

The applications of the moving average procedure and related data interpretation are demonstrated in detail in the examples below. In this section, the description of the results of the statistical analysis is supplemented by a graphical representation, which in this case turned out to be more informative than the corresponding confidence intervals and p-values.

3 Examples

3.1 Example 1

The distributions of $Y_i(X = 0)$ and $Y_i(X = 1)$ are not significantly different from normal (symmetric) distributions.

Let us illustrate the idea and application of the moving average to independent observations using as an example a study on the dependence of human body mass index Y on smoking status X. The study involved 375 male students aged from 16 to 17, studying at secondary vocational schools in the city of Chelyabinsk (Russia) (Varaksin et al. (2022)). The young men were examined for anthropometric characteristics and smoking status. Out of the anthropometric data, we took for the purposes of this example the body mass index ($BMI = weight/height^2$, where weight is determined in kg and height in meters), which is a continuous quantitative variable. The dichotomous variable X in this case is represented by the variable "Smoking" assuming two values $X = 0$ (a non-smoker) and $X = 1$ (a smoker). For the analysis, we selected 226 young men who did not, and had never, smoked, and 149 young men who had smoked for more than two preceding years.

Thus, we obtained two samples of values for the quantitative variable, $Y_i(X = 0)$ and $Y_i(X = 1)$. The distributions of Y (i.e. BMI in our case) in the two groups were not different from normal distribution (Kolmogorov-Smirnov test, $p > 0.37$; neither did we find any significant differences in Y variance between the groups (F-test, $p = 0.152$); hence, we have the normal (symmetric) distribution for outcome Y.

The t-test enabled comparing the average values of the quantitative variable in the two samples, $\langle Y(X = 0)\rangle$ and $\langle Y(X = 1)\rangle$ (Figure 1). The application of the t-test provided the following results: $\langle BMI(X = 0)\rangle = 21.46$ kg/m^2; $\langle BMI(X = 1)\rangle = 21.02$ kg/m^2. Thus, the average BMI value in the group of smokers proved to be less than in the non-smoker group, but the differences in average value are very small and statistically insignificant ($p < 0.156$), even though the number of observations is fairly large (375 young men).

Let us apply the moving average procedure to the above set of observations (Varaksin et al. (2022)). After applying the moving average procedure to the *BMI* variable, with a window size of $n_w = 20$, we obtain 356 strata (the number of strata is equal to $n = 375$ observations minus $n_w = 20$ plus 1); for each stratum, we compute the average *BMI* and the average value of the ex-dichotomous variable "Smoking". Now the average value of the variable "Smoking" represents the proportion of smokers in a stratum; by multiplying the proportion of smokers by 100 we obtain the percentage of smokers. The result is shown in Figure 2.

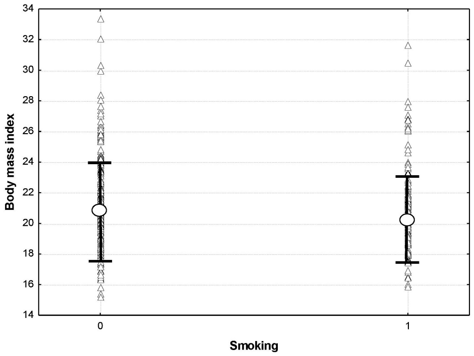

Figure 1: Primary data with the t-test applied. The smaller open circles show individual *BMI* values (in kg/m^2) for each of the 375 young men; the larger circles show average *BMI* values, and the whiskers present 95% confidence intervals for the average *BMI* values in the groups of non-smokers (0) and smokers (1)

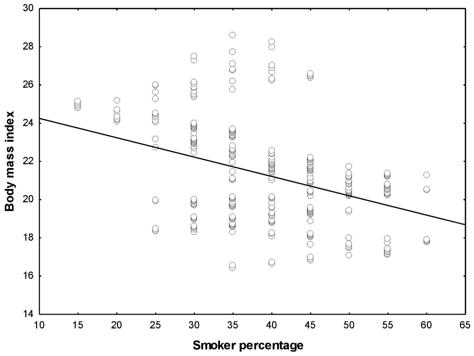

Figure 2: Scatterplot for the moving average data: the abscissa plots the average values of the dichotomous variable "Smoking" obtained by the moving average procedure; the ordinate shows the average *BMI* values in 356 strata

In the "ideal" case, the percentage of smokers should vary from 0 to 100. In our data, we found that for $n_{\mathrm{w}} = 20$ the percentage of smokers in the actual strata varied from 15 to 60 with a step of 5 (the total number of strata was 10), while with the t-test applied, the percentage of smokers assumed only two values: 0 and 100. Figure 2 shows an explicit decrease in average *BMI* values as the percentage of smokers increases. The straight line in Figure 2 is a linear regression line, described by the equation:

$$BMI = 25.26 - 0.10 \cdot Smoker\ percentage.$$

The correlation coefficient $r = -0.404$ between *BMI* and the percentage of smokers is statistically highly significant ($p < 0.000001$), which means there is a significant statistical relationship between *BMI* and smoking.

Figure 2 highlights also the distribution of average *BMI* values in different groups by smoking. Thus, for the percentages of smokers equal to 15 and 20, we can see a narrow scattering of *BMI* values (from 24 to 25 kg/m^2); for the smoker percentages of 35 and 40, the *BMI* values are widely scattered (approximately from 16 to 29 kg/m^2); finally, as the percentage of smokers increases to 50 and higher, *BMI* data scattering decreases again (from 17 to 22 kg/m^2).

The regression line in Figure 2 is a very simple interpolation of *BMI* values and smoker percentages. For a more robust approximation of the relationship between *BMI* and smoker percentage, we can take advantage of the fact that the *BMI* values for the groups of smokers and non-smokers follow the normal distribution law with equal variances. It is known that in this case, the relationship between smoker percentage X and *BMI* is given by (Hosmer et al. (2013), Shoukri et al. (1999))

$$log\left(\frac{X}{100-X}\right) = a_0 + a_1 \cdot BMI,$$

where a_0 and a_1 are constants.

By inverting this relationship, we have

$$BMI = b_0 + b_1 \cdot log\left(\frac{X}{100-X}\right). \tag{1}$$

The regression line (1) with coefficients $b_0 = 20.09$ and $b_1 = -2.706$, obtained by applying the least squares technique, is shown in Figure 3. This figure also shows the *BMI* average values and confidence intervals in the groups of different smoker percentages. The average *BMI* values in the groups are seen to decrease from 24.9 kg/m^2 in the first group with the proportion of smokers of 15% to 18.8 kg/m^2 in the last group with 60% smokers. This range of average *BMI* values (from 18.8 kg/m^2 to 24.9 kg/m^2) is considerable in terms of adolescent hygiene since it covers a wide span from the 25th to the 75th percentile of the WHO growth standards for young males aged 17 (WHO (2006)).

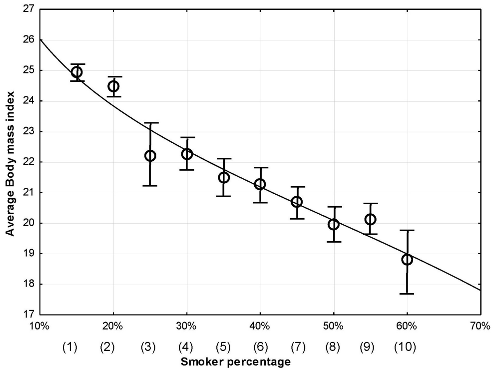

Figure 3: *BMI* as a function of smoking. Circles represent average *BMI* in the groups by smoker percentage, whiskers are 95% confidence intervals, and the solid line is the regression line described by equation (1)

A comparison between Figure 3 and Figure 1 reveals clearly a significant relationship between *BMI* and smoking when the moving average is used. Note that both figures display the same: the axes plot *BMI* and smoking, and the circles and whiskers in both cases show average *BMI* values and related confidence intervals. For estimating the statistical significance of the differences between the average *BMI* values in the groups, we used a multiple comparison procedure as shown in Figure 3, which replaces t-test comparison if the number of groups to be compared becomes greater than two. The multiple comparison procedure (Scheffe test in analysis of variance (Afifi et al. (1979))) reveals the following significant differences ($p < 0.05$) between the average *BMI* values of the groups in Figure 3: the first group differs from groups 7-10; the second group, from groups 7-10; and the fourth group, from groups 8-10. Thus, for the division of *BMI* values into 10 groups instead of the initial division into two groups smokers and non-smokers, we observe significant differences in average *BMI* values between the groups by smoker percentage (even applying a multiple comparison procedure which increases p-values with an increase in the number of groups). Thus, in contrast to the t-test, the moving average helps reveal a statistically and hygienically significant effect of smoking on adolescent *BMI*.

Because this is a methodological paper, we do not discuss in detail the causes behind reduced *BMI* in smokers in this particular case, which is a question for epidemiologists and hygienists rather than biostatisticians. We should emphasize, however, that the moving average procedure may be used for identifying possible causes of *BMI* reduction. To this end, after dividing *BMI* values into strata by the moving average technique, it would be necessary to compute average values not only for *BMI* and smoking but also for various other factors that are likely to affect *BMI*. We performed such calculations for our sample in our recent work (Varaksin et al. (2022)) and discovered significant relationships between smoking and some foods; these relationships are likely to explain the results obtained (Figure 3).

3.2 Example 2

The distributions of $Y_{\mathrm{i}}(X = 0)$ and $Y_{\mathrm{i}}(X = 1)$ are strongly asymmetric.

A lot of variables in biostatistics display an asymmetric distribution, for example, log-normal distribution as reported by Shoukri et al. (1999). Let us use this example to illustrate how the moving average could be used for finding statistical relationships between the quantitative variable Y and the dichotomous X in a case where Y is asymmetrically distributed. In this asymmetric case, the possible analytical options are Student's test for a large number of observations; data transformation towards a less asymmetric distribution (e.g., taking the logarithm of lognormal distribution); or use non-parametric methods.

The example is taken from a study of heart rate variability data for more than 1500 industrial workers examined in the Urals (Russia) in the course of preventative occupational health checks (Maslakova et al. (2021)). Out of this array of observations, we selected 203 males who did not, and had never smoked, and 456 males who had smoked for more than 10 years; all men were over 35 years of age.

For these men, we investigated the relationship between smoking status (dichotomous $X = 0, 1$) and Stress Index (quantitative variable) or *SI* (baseline test for heart rate variability)

$$SI = \frac{AMo}{2 \cdot Mo \cdot Var}, \tag{2}$$

where *Mo* is the Mode (the most frequently observed cardiointerval), *AMo* is the Mode amplitude (the number of cardiointervals corresponding to the value of the Mode, as a percentage of the sample size), and *Var* is the Variable Range (difference between the maximal and minimal values of cardiointervals).

Figure 4 shows *Stress Index* histograms in the groups of non-smokers and smokers. The histograms confirm the well-known fact that the *Stress Index* distribution is very asymmetric (Baevsky et al. (2017)). Testing showed that the *SI* distribution in the non-smoker group was lognormal while in the smoker group it was statistically different from lognormal (just an asymmetric distribution).

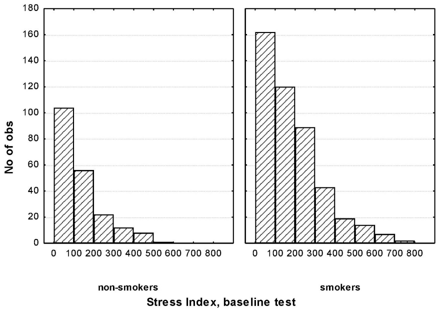

Figure 4: Histograms for the distribution of *Stress Index* values in the groups of male non-smokers and smokers

The statistical significance of the differences in *Stress Index* between the groups of smokers and non-smokers may be estimated by non-parametric statistics. For instance, the Mann-Whitney U test shows differences in *Stress Index* rank between the groups of smokers and non-smokers, and we found that the groups were different significantly (p = 0.000001). In any case (whether significant differences are present or not), non-parametric tests do not help reveal the magnitude of differences in *Stress Index* itself between the groups, showing just differences in rank; as reported by Hettmansperger (1984).

Now let us apply the moving average procedure to the above epidemiological case in relation to *Stress Index*, the window being $n_{\mathrm{w}} = 10$ observations. Figure 5 shows a scatterplot for *Stress Index* averages (the ordinate) and smoker percentages (the abscissa) in the strata; the number of strata is 203 + 456 males minus $n_{\mathrm{w}} = 10$ plus 1 and is equal to 650.

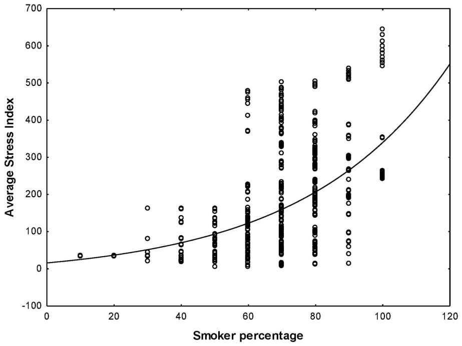

Figure 5: Scatterplot for the moving average data: the abscissa plots the average values of the dichotomous variable "Smoking" obtained by the moving average procedure; the ordinate shows the average *SI* values in 650 strata. The curve line is an exponential regression line

Figure 5 highlights the distribution of *Stress Index* averages in different groups by smoking. Normally (in the healthy organism), the *Stress Index* should be within the interval from 50 to 150 units (*SI* should not be less than 50 units or more than 150 units) (Bayevsky et al. (2001)). Since smoking is a well-known cardiovascular risk factor, an increase in the percentage of smokers should reveal itself in a greater value of *SI* (over 150), which is what we see in Figure 5. Also, as the smoker percentage increases, the proportion of low *SI* values should also increase (under 50), which is not observed in Figure 5. A similar behavior (absence of any large proportion of low *SI* values in the presence of the risk factor) was observed when the cardiovascular risk factor considered was radiation (Konstantinova et al. (2017)).

Figure 5 shows that the relationship between *SI* and smoker percentage is clearly non-linear (in contrast to Figure 2). For approximating this relationship, we used the exponential model of the following form

$$SI_{Av} = b_1 + exp(b_2 + b_3 \cdot Smoker\ percentage) \tag{3}$$

If we use the loss function as least squares, the model coefficients are as follows: $b_1 = 3.372$; $b_2 = 3.067$; $b_3 = 0.02783$ (all coefficients are statistically significant, $p < 0.05$). For this model, the determination coefficient $R^2 = 0.258$.

The exponential pattern of the relationship between *SI* and smoking is more explicit in Figure 6, which shows all *SI* values from Figure 5 averaged over the group with the same smoker percentage. The average *SI* values in the groups are seen to increase from 35 in the first two groups with the smoker percentages of 10% and 20% to 360 in the last group with 100% smokers.

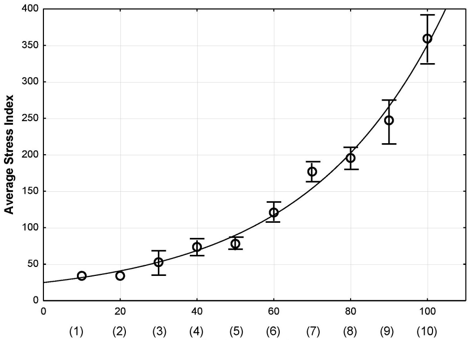

Figure 6: *SI* as a function of smoking. Circles are average *SI* values in the groups by smoker percentage; whiskers are 95% confidence intervals; the curve line is the line of exponential regression described by equation (3)

A comparison of Figure 5 with the similar Figure 2 reveals a difference between the results obtained by applying the moving average to variables Y with symmetric and asymmetric distributions: in the first case, the scatter chart looks like a two-dimensional normal distribution while in the second case the diagram is explicitly skewed (triangular). In Figure 3, the average values of *BMI* fit in well with the line describing the regression function in Figure 2. Figure 6 shows clearly that the average values of *SI* are non-linearly related to smoker percentage X (the relationship between the average values of *SI* and X tends to be exponential). Thus, the approach based on the moving average enables the assumed shape of the statistical relationship between the quantitative response Y and the dichotomous predictor X to be determined, which appears to be markedly different for outcomes Y with symmetric and asymmetric distributions.

The considered examples allowed us to show the features of the proposed approach. Strengths of the study: (i) our work has added a new tool to the set of methods for establishing a relationship between a quantitative outcome, and a dichotomous predictor; (ii) the proposed method allows us to visualize the hidden structure of the relationship between variables. The weaknesses of the moving average method study include the fact that the minimum value of n (number of cases) is not uniquely determined; however, it has been empirically established that for a selected window size of $n_{\mathrm{w}} = 10$, the sample size should not be less than 20.

4 Concluding Remarks

The conventional parametric method for evaluating relationships between quantitative outcome Y and dichotomous predictor variable X is the t-test or its variants.

We proposed a new moving-average-based technique which does not replace the existing methods of analysis; rather, it expands the range of methods available for analyzing relationships between variables Y and X: now it makes it possible to use the methods of linear regression and analysis of variance. An analysis of actual epidemiological data demonstrates that the new technique enables one (in some cases) to reveal different statistical relationships between Y and X for both symmetric and strongly asymmetric distributions of Y.

We showed that the methods of visualization of multidimensional data and the approaches based on them are promising for the analysis and visual presentation of experimental numerical results from the field of biology and medicine.

References

Afifi, A. A. and Azen, S. P. (1979), *Statistical analysis: a computer oriented approach*. New York: Academic Press.

Baevsky, R. M. and Chernikova, A. G. (2017), *Heart rate variability analysis: physiological foundations and main methods*. Cardiometry, **10**, 66–76. doi: 10.12710/cardiometry.2017.10.6676.

Bayevsky, R. M. and Ivanov, G. G. (2001), *Cardiac Rhythm Variability: the Theoretical Aspects and the Opportunities of Clinical Application (Lecture)*. Ultrasound and Functional Diagnostics, **3**, 108–127.

Büning, H. (2002), *Robustness and power of modified Lepage, Kolmogorov-Smirnov and Crame´r-von Mises two-sample tests*. Journal of applied statistics, **29**(6), 907-924. doi: 10.1080/02664760220136212.

Gerald, B. and Patson, T. F. (2021), *Parametric and Nonparametric Tests: A Brief Review*. International Journal of Statistical Distributions and Applications, **7**(3), 78-82. doi: 10.11648/j.ijsd.20210703.12

Hazra, A. and Gogtay, N. (2016), *Biostatistics series module 3: Comparing groups: Numerical variables*. Indian journal of dermatology, **61**(3), 251-260. doi: 10.4103/0019-5154.182416.

Hettmansperger, T. P. (1984), *Statistical inference based on ranks*. Nashville: John Wiley & Sons.

Hosmer, D. W., Lemeshow, S. and Sturdivant, R. X. (2013), *Applied Logistic Regression*. Nashville: John Wiley & Sons.

Kim, T.K. and Park, J.H. (2019), *More about the basic assumptions of t-test: normality and sample size*. Korean Journal Anesthesiol, **72**(4), 331-335. doi: 10.4097/kja.d.18.00292.

Konstantinova, E. et al. (2017), *Imbalances of the autonomic nervous system in people living in radioactive contaminated territories: radioactivity and vegetative imbalance*. International Journal of Radiation Research, **15**(3), 317-320.

Krispin, R. (2019), *Hands-on time series analysis with R: perform time series analysis and forecasting using R*. Birmingham Mumbai: Packt Publishing.

Le Cessie, S., Goeman, J. J. and Dekkers, O. M. (2020), *Who is afraid of non-normal data? Choosing between parametric and non-parametric tests*. European journal of endocrinology, **182**(2), E1-E3. doi: 10.1530/EJE-19-0922.

Maslakova, T. A. et al. (2021), *Assessment of the state of the cardiovascular system in metallurgical workers depending on the length of service in harmful working conditions, behavioral and metabolic risks*. Ecological Systems and Devices, **10**, 18-28.

Mosteller, F. and Tukey, J. W. (1977), *Data analysis and regression: A second course in statistics*. New Jersey: Upper Saddle River.

Nuijten, M. B., Hartgerink, C. H., van Assen, M. A., Epskamp, S. and Wicherts, J. M. (2016), *The prevalence of statistical reporting errors in psychology (1985–2013)*. Behavior Research Methods, **48**(4), 1205–1226. doi: 10.3758/s13428-015-0664-2.

Shoukri, M. M. and Pause, C. A. (1999), *Statistical methods for health sciences*. Florida: CRC Press.

Student (1908), *The Probable Error of a Mean*. Biometrika, **6**(1), 1-25. doi: 10.2307/2331554.

Tukey, J. W. (1977), *Exploratory Data Analysis*. New Jersey: Upper Saddle River.

Varaksin, A. N. et al. (2022), *An analysis of the links between smoking and BMI in adolescents: A moving average approach to establishing the statistical relationship between quantitative and dichotomous variables*. Children, **9**(2), 220-233. doi: 10.3390/children9020220.

Walker, G. A. and Shostak, J. (2019), *Common statistical methods for clinical research with SAS examples*. Cary, NC: SAS Institute.

World Health Organization (2006). WHO child growth standards: length/height-for-age, weight-for-age, weight-for-length, weight-for-height and body mass index-for-age: methods and development. https://www.who.int/publications/i/item/924154693X.

Yaffee, R. A. and Mcgee, M. (2000), *Introduction to time series analysis and forecasting: with applications in SAS and SPSS*. San Diego: Academic Press.

Yaffee, R. A. (2007), *Stata 10 (Time Series and Forecasting)*. Journal of Statistical Software, **23**.

Zimmerman, D. W. and Zumbo, B. D. (1992), *Parametric alternatives to the student T test under violation of normality and homogeneity of variance*. Perceptual and motor skills, **74**(3), 835-844. doi: 10.2466/pms.1992.74.3.835.