ISSN (p) 2303-4890, **ISSN (o) 1986–518X** *ISTRAŽIVANJE MATEMATIČKOG OBRAZOVANJA*
http://www.imvibl.org/dmbl/meso/imo/imo2.htm
Vol. VII (2015), Broj 12, 37-46
Original research paper

SOME COMMON MISTAKES OF DATA ANALYSIS, THEIR INTERPRETATION, AND PRESENTATION IN BIOMEDICAL SCIENCES

Selman Repišti[1]

Abstract: The main aim of this paper is to present and discuss several commonly made mistakes regarding data analyses, interpretation of results as well as how they are reported in scientific papers. The provided examples are linked to the biomedical sciences, in order to put this presentation and discussion in the context of one of the fields where contemporary research is not possible without deep understanding of statistical procedures. There were discussed the following topics/issues: both graphical and tabular data presentation, the process of detecting outliers, data cleansing procedures, the chi-squared test, calculation of descriptive statistical values for non-normally distributed data, reporting risk and odds ratios, counducting t-tests, reporting confidence intervals and measures of size effect, calculating coefficients of correlation, and drawing conclusions based on the presented results.

Keywords: statistical mistakes, data analysis, reporting results, biomedical statistics.

Sažetak: Glavni cilj ovog rada je prikaz i rasprava o nekoliko uobičajenih grešaka prilikom analize podataka, interpretacije rezultata, te načina na koje se oni prikazuju u naučnim člancima. Izloženi primjeri su vezani za biomedicinske nauke, kako bi se ovaj prikaz i diskusija smjestili u kontekst jednog od područja u kojem savremeno istraživanje nije moguće bez dubokog razumijevanja statističkih procedura. Ovdje su obrađene sljedeće teme/problemi: grafičko i tabelarno prikazivanje podataka, proces detekcije ekstremnih vrijednosti, načini "čišćenja" podataka, hi-kvadrat test, računanje deskriptivnih statističkih vrijednosti za podatke koji nisu distribuirani po normalnoj raspodjeli, izvještavanje o količnicima rizika i šansi, provođenje t-testova, izvještavanje o intervalima povjerenja i mjerama veličine efekta, računanje koeficijenata korelacije i izvođenje zaključaka na osnovu izloženih rezultata.

Ključne riječi: statističke pogreške, analiza podataka, navođenje rezultata, biomedicinska statistika.
Mathematics Subject Classification (2010): 97B40, 97D70, 97K40, 97K80, 97M60
ZDM Subject Classification (2010): B40, D70, K40, K80, M60

Introduction

Nowadays, conducting statistical analyses and drawing conclusions from them is the essential part of the process of writing scientific articles. Medicine and psychology are the fields where statistical analyses are inevitable. Furthermore, variables included in research in these two fields have very complex interrelationships. Because of that, researchers in medicine and psychology usually have to apply a wide range of different statistical procedures. These procedures are the following [10, 18, 19]: univariate techniques (such as descriptive statistics), bivariate techniques (e.g. correlation/ association analyses) and multivariate techniques (multiple linear/nonlinear regression analysis; ROC curve analysis, logistic regression analysis, multivariate analysis of variance (MANOVA) etc.).

However, some of the authors/researchers are not very familiar with the logic which lies behind the aforementioned procedures as well as they have problems while interpreting output provided by different statistical softwares (e.g. SPSS, R, STATA, SAS, MS Excel...). They often do not know clearly what information should be reported in their papers. In general, there are four types of mistakes in reporting their results:

[1] Sarajevo, Bosnia and Herzegovina, e-mail: selman9r@yahoo.com

1) The omission of some important indicators (such as confidence intervals or effect size).
2) Reporting too many pieces of information (e.g. calculating the means and standard deviations for every variable which is not of a main interest in the study, reporting too many decimals in the case of test values...).
3) Incorrectly perfoming statistical procedures which are not suitable for a particular set of data (e.g. calculating Pearson's product-moment correlation coefficient for data given on the ordinal scale, instead of using Spearman's rho coefficient of correlation).
4) Mispresentation of the results of their study (e.g. using bar plots to present arithmetic means graphically).

Strasak, Zaman, Pfeiffer, Göbel, and Ulmer [21] identified the following four types of statistical errors and deficiencies related to data analysis: use of wrong statistical tests, inflation of the Type I error, several errors related to Student's t-test, and typical errors with chi-squared test. These authors also listed some errors which can be made in documentation of statistical methods applied as well as statistical deficiencies in the design of a study. Lang [13] described and discussed 20 statistical errors which can be found in biomedical research articles. Some of them are: reporting measurements with three or more digits (i.e. with unnecessary precision), using descriptive statistics incorrectly, reporting only p-values which are often misinterpreted, using tables and figures in a manner different from the basic principle of communication – clarity (i.e. sometimes readers can find tables and figures very difficult to understand), and confusing statistical significance with clinical significance (i.e. importance). Tenjović and Smederevac [22] highlighted the importance of reporting confidence intervals and effect size, rather than p-values only. These kinds of mistakes may also occur as a results of the following [6]: poorly defined aim of a study, mistreatment of outliers, extrapolation issues (e.g. making broader conclusions based on limited information, for example, on a small set of data or results), casually sampling (undersampling – removing common cases from the analysis and oversampling – duplicating rare/odd cases), etc. These issues can be classified as methodological ones, however, they are related to the statistical data processing. Moreover, bad methodology often implies bad statistics and as a result of this – invalid conclusions and generalizations.

Festing and Altman [7] discussed on what statistical procedures are more suitable for the analysis of data. They underlined the importance of using parametric, rather than non-parametric techniques, because the first ones are more powerful (i.e. the power of these tests is higher). Normality allows drawing reliable conclusions from various statistical estimates [8]. If the data do not follow normal distribution, we can transform them in order to be distributed as close as the Gaussian curve). If these transformations do not produce the appropriate result, we can use non-parametric techniques (such as Mann-Whitney test, Kruskall-Wallis test, Kendall's tau coefficient of correlation etc.). Han and Camber [9] stressed that the measures of central tendency have to be representative of a paticular dataset as well as robust to extreme values. Osborne and Waters [15], for example, listed and explained four assumptions of multiple regression that researchers have to test before conducting this kind of analysis: checking for normality of variables; examining if the relationship(s) between independent variable(s) and dependent variable is linear, checking reliability of variables, and determining if the assumption of homoscedasticity was met. Lots of researchers skip these steps and conduct linear regression analysis right away.

For these and many other reasons, it is very important to inform researchers on the appropriate way of conducting data analyses as well as how to interpret the output obtained on the basis of the analyses applied. This is the main scope of this paper. The other aim of this review paper is to help researchers and students in their effort to understand, apply and present their results. *Hence, this article should also be seen as educational material for young researchers and those who want to refresh their knowledge in the biomedical statistics.*

Common mistakes of data analysis

Some researchers do not pay attention to the *data cleaning (cleansing) process*. According to Pallant [17], there are three steps of the data cleansing process: checking for errors (finding scores of variables that are out of a particular range), finding where (in our dataset) this error exactly occurred, and correcting it.

Thus, before any analysis, we have to ensure that our database *does not contain incorrect numbers* (e.g. there is a number of 33 instead of 3, which was mistakenly entered twice). We can check it by giving our program the command to display frequencies of data. For example, if our measuring scale consists of five points (from 1 to 5), it is not possible that our participant scored 7 or 55 points on it. Therefore, we have to find this value(s) and make corrections. If we do not correct them, we will not be able to obtain the real value of arithmetic mean or standard deviation. For example, we asked five participants (patients) to estimate the quality of relationship with their physicians. For this purpose, we used Likert's five-point scale. We entered results in a statistical software and we listed them as follows:

4 2 3 4 55

The average value (the mean) of this dataset was $M = 13.6$, whereas the standard deviation was SD = 23.16. Then, we realized that we mistakenly entered the result "5" twice. We corrected it and now we have the following situation: M = 3.6 and SD = 1.14. We saw that just one value can dramatically change measures of centrality and dispersion.

We can also *skip a value* and *enter it in the next variable column*:

4 2 . 34 5

It is obvious that there is a number "3" instead of dot (4, 2, 3, 4, 5). Hence, we have to check our database and if it looks fine, we could continue with our analysis.

In some cases, we can have valid results (values) which are far away from the mean of the data. These results are called *outliers* or extreme values. They influence parametric measures of central tendency and variability. We can exclude them, by examining boxplots. Boxplots are diagrams which represent the five important numbers (or points) of a dataset [e.g. 23, 1, 14]: median (i.e. 50^{th} percentile), 25^{th} percentile, 75^{th} percentile, minimum value and maximum value. If some values fall below Q1 – 1.5*IQR (this is called lower fence) or above Q3 + 1.5*IQR (i.e. above the upper fence; IQR is interquartile range, calculated by subtracting result that corresponds to 75^{th} percentile from the result that corresponds to 25^{th} percentile), they are considered as outliers. If we want to retain them, we have to use non-parametric procedures, because the median (C) is a more robust measure of central tendency, compared to the mean (M). We can also use the interquartile range (IQR, which is the difference between results that correspond to the 75^{th} and 25^{th} percentile), instead of the standard deviation (SD).

When we conduct statistical tests, we must be aware of the following thing: *every statistical procedure has some conditions to be met before we apply that procedure.* For example, if we want to compare two groups of participants by applying t-test, we have to ensure that our variables follow the normal distribution and that variances in these two groups are not significantly different. We cannot compare 15 males with 75 females, because there is a huge difference between the sample sizes. Nonetheless, we can compare samples consisted of e.g. 120 and 135 participants or 200 and 235 participants, because the relative differences between these pairs of samples are not as big as in the previous example. If the variances differ significantly, then we should use the value of t-test and its *p*-value that are given in the part of the output, called "the equality of variances not assumed" (this example is based on SPSS output). If the data distribution differs from the normal curve, we can use the Mann-Whitney U test, which compares the ranks of results.

When researchers use t-tests, they have to know which type of t-test is appropriate for a particular data. If we want to determine differences in average results between two groups (e.g. experimental and control group), we use t-test for independent samples. However, when comparing the effects of several treatments on the same group of participants, t-test for paired-samples is used.

The next issue concerns conducting lots of t-tests in order to examine whether there are statistically significant differences. If we use lots of tests (i.e. repeat this procedure many times), we will increase the probability of accepting alternative hypothesis which could be, in fact, wrong – this is called the error Type 1 [20]. We recommend that there should be applied the *Bonferroni correction*. The logic which lies behind this method is very simple. If we want to conduct five t-tests (or e.g. five ANOVAs), we have to divide the significance level by five. This commonly used significance level is $p < .05$ (if the

obtained significance value is lower, e.g. $p = .02$, we can refuse our null-hypothesis and accept the alternative one; if it is higher e.g. $p = .12$, then we have to accept/confirm our null hypothesis). Therefore, the new significance level will be $p < .01$. Then, we will compare our obtained *p*-values with the new one and decide to accept of refuse our null hypothesis. If we want to conduct seven analysis of the same type, we should divide the significance level by 7, and so on.

The next issue is correlation. The Pearson's coefficient of correlation is suitable for data which are obtained from the interval measurement scale. In addition, this coefficient requires normally distributed results and linear relationship between variables. If these conditions are not met, we will have to use non-parametric coefficients of correlation (e.g. Spearman rank-order correlation coefficient). Suppose we want to correlate the stages of breast cancer with perceived social support in female patients. Stage of breast cancer is a variable measured at the nominal level. However, it can be considered as ordinal variable, because e.g. the fourth (IV) stage is more severe than the third stage (III) is. On the other side, perceived social support was an interval variable (it can be measured by different psychological scales, where the intervals between scale points are the same). In this example, we cannot use Pearson's coefficient of correlation, because one of our variables is not measured at the interval level. We can calculate Spearman's rho (ρ or r_s) coefficient of correlation and determine the size of this relationship.

In correlational studies, it is also important to calculate the coefficient of determination (r^2). When we multiply it by 100, we get the percentage of common (shared) variance between variables.

In the case of conducting chi-squared test, we should be aware of some important procedures. First, we have to code categories, variable levels or answers correctly. It has to be done in order to avoid ambiguities or misunderstandings. For example: yes – 1, no – 0; alive – 1, deceased – 0; positive – 1, negative – 0 etc.

If we have a 2 x 2 contingency table (e.g. when calculating association between the number of patients who have/have not treated surgically and the number of those who survived/not survived) we should use Yates's correction for continuity. We also use it for 1 x 2 tables. This correction has to be done for small data in order to prevent overestimation of statistical significance. Additionally, this is suitable for expected cell values less than 5. However, because some statisticians (e.g. [3]) disagree with its usage, the Fisher's exact test showed to be the best solution for small samples.

Finally, the issue of moderator variable(s) is very important. If we have information (e.g. from the previous studies in a particular field) that some variable(s) can affect the size and/or strength of relationship (or association) between our main variables, we must include it/them in the analysis. Those variables are, for example, age group, gender, social class, etc [2]. We can divide our sample in two subsamples – males and females, and separately calculate coefficients of correlation or association. Or we can include variable "gender" in our main analysis and determine its interaction with other variables.

Common mistakes made while interpreting results

The fundamental error in the interpretation of results is drawing conclusions, when statistical tests are not conducted. Somebody can conclude that $M_1 = 40.9$ is significantly higher than $M_2 = 38.5$. This could be true, but we must apply inferential statistics first (e.g. *t*-test). If samples are very small, the probability that this difference is statistically significant is small. In addition, if standard deviations are high, this can also lead to the non-significant result. For example, if $SD_1 = 8$, $SD_2 = 7$, and $n_1 = n_2 = 90$, we will get $p = .034$, which is a significant result ($p < .05$). However, if $SD_1 = 10$, $SD_2 = 8$, and $n_1 = n_2 = 45$, we will get $p = .212$, which is a non-significant result ($p > .05$). Therefore, every researcher should provide complete information on these values as well as perform adequate statistical tests. *Without applying inferential statistical procedures, there is no reason to make conclusions.*

Further, all researchers should understand the *scientific notation*. Suppose that output contains the following information: "2.04E+4" or "1.56e-05". Some of the researchers think like this: "E (*e*) probably stands for the Euler's number ($e \approx 2.72$)" and they could multiply 2.04 by the Euler's number, and add 4 (the first example). However, these are examples of writing results in scientific notation, where

2.04E+4 is equal to $2.04*10^4$ and 1.56e-05 stands for $1.56*10^{-5}$. It is clear that incorrectly interpreted results can lead to false conclusions.

Some scholars who are not familiar with statistics, but want to perform some statistical procedures, can think that the standard error of the mean is standard deviation. In fact, standard deviation is the square root of variance, while the standard error of the mean (*SE*, *SEM*, or SE_M) is the standard deviation of estimates of the true (population) mean.

Another type of mistake can be made while interpreting contingency tables. Suppose that we obtained the results shown in Table 1. We consider two variables: treatment assignment (yes/no) and the number of recovered and non-recovered persons. Some researchers interpret these numbers in the following way: the treatment showed to be useful and successful because there are more recovered people (n = 30) in the group with treatment than in the group which did not receive the treatment (n = 20). However, if we calculate the percentage of those cured in the group with the treatment and compare it to the percentage of the cured people in the second group, we will realize that these percentages are the same (i.e. both of them equal to 60%). We can also prove that χ^2 –statistic is equal to zero. Therefore, there are no differences in the rates of recovery and the treatment is likely useless.

Table 1. Example for interpreting numbers in contingency tables (determining the effects of treatment)

	Groups	
Treatment	Cured	Not cured
Yes	30	20
No	15	10

The next example deals with making conclusions, based on the statistically significant result of chi-squared test. Suppose we want to examine whether body mass index (BMI) is associated with physical activity. Based on the values of their BMI, participants were grouped into four distinct categories, whereas physical activity is given as the number of hours per week spent in doing exercises, running, climbing, etc. (Table 2).

Table 2. Example for concluding from the significant chi-squared value (BMI and physical activity)

	Physical activity (hours per week)			
BMI	< 1 hour	1 or 2 hours	3 or 4 hours	> 4 hours
< 18.5	15	35	30	12
18.5 – 24.9	16	60	15	22
25 – 29.9	45	18	24	27
≥ 30	40	30	48	10

If we calculate χ^2–statistic, based on the numbers presented in the Table 2, we will obtain the following result: $\chi^2 = 74.89$, $df = 9$, $p < .001$. Therefore, we got a statistically significant result. A researcher who was not well-trained in statistics can conclude: the higher value of BMI, the lower physical activity. But, is it a valid interpretation of our results? Certainly not, because this researcher did not examine the percentage values that can be calculated regarding the numbers presented in the Table 2. We expect that underweight participants (BMI < 18.5) spend more hours (e.g. more than four hours) in physical activities than overweight participants (25 ≤ BMI ≤ 29.9) do. However, 13.04% (n = 12 out of 92) of those who were underweight spend in it more than four hours, whereas 23.68% (n = 27 out of 114) of those who were overweight spend it so. Thus, in this case, owerweight participants dedicate larger amount of time to physical activity (compared to underweight participants). Next, 32.61% of the underweight participants (BMI < 18.5) and 37.50% of the obese participants (BMI ≥ 30) spent 3–4 hours per week on their physical activity. Hence, in the case of 3-to-4-hour-activity, the percentage of obese

participants is higher than the percentage of underweight participants. On the other hand, there was a higher percentage of underweight and normal participants ($18.5 \leq BMI \leq 24.9$) who spend one or two hours per week in physical activity, than those who belong to one of the following groups – "overweight" and "obese".

Hence, the significant value of the chi-squared statistic is not enough for concluding about the association between two variables. We have to inspect table values and, even more important, percentages. Significant χ^2–statistic implies only that some observed and expected frequencies differ significantly. In this case, it does not tell us which pairs of the frequencies differ significantly and does not lead to an unambiguous conclusion!

Common mistakes in the presentation of results

The sample mean is usually labelled as *M*, whereas sample standard deviation is usually denoted as *SD* or *s*. These values are called estimates (or statistics). On the other hand, there are also the parameters – the true (population) mean (labelled as μ) and the population standard deviation (σ). Some researchers use these Greek letters to label sample values (i.e. estimates), which is an incorrect way to present them in scientific papers.

While presenting their data, lots of researchers omit degrees of freedom (*df*). According to Eisenhauer [5], degrees of freedom are usually equal to the number of independent results (i.e. sample size or *N*) minus the number of values (somewhere called parameters) used in the estimation of a particular parameter (i.e. the value of our interest).

For example, in the correlation analysis as well as in the t-test for independent samples, $df = N - 2$ (where *N* is sample size, or it equals to the sum of participants in two groups that we want to compare). In paired-samples t-test, $df = N - 1$ (where *N* is just the number of our participants). In the chi-squared test with only one categorical variable (e.g. participants' race), the degrees of freedom are equal to the number of categories minus one ($df = k - 1$). When tables include two variables with several categories, the number of degrees of freedom is equal to $r - 1$ times $c - 1$ (where *r* and c stand for the number of rows and columns, respectively). In the analysis of variance (ANOVA), while testing the impact of two independent variables, their degrees of freedom are equal to their number of categories minus one. For example, gender comprises two categories – "males" and "females", therefore, the number of its degrees of freedom is $df = 1$. If we test for the effect of interaction of two categorical variables (e.g. the first one is *A* with *a* categories and the second one is B including *b* categories), the number of degrees of freedom is $a - 1$ times $b - 1$. As we can see, the number of degrees of freedom can be calculated by using simple procedures and there should be no excuse to exclude it while reporting results.

Next, it is *not enough to report only the p-value, if we want to refuse or accept our null hypothesis.* We have to report confidence interval (i.e. its lower and upper limit), as well. For t-test, the confidence interval is equal to $M_1 - M_2 \pm t_{CL} * SE_{M1-M2}$ (i.e. the difference between two means plus/minus t-value for the desired confidence level multiplied by the standard error of means' difference).

When we consider the relative risk (risk ratio) and odds ratio, we will provide formulas for calculating their confidence intervals (referring to Table 3).

Table 3. Reference table for calculating RR and OR confidence intervals

	Cases	Controls	*Total*
Exposed	a	b	a+b
Not exposed	c	d	c+d
Total	*a+c*	*b+d*	N

The value of risk ratio is calculated by using the following formula [16]:

$$Ln(RR) \pm z_{\alpha} * sqrt(1/a + 1/c - 1/(a+b) - 1/(c+d)),$$

where *Ln(RR)* is a natural logarithm of the value of risk ratio, *sqrt* is the square root of this sum, z_α is a number that represents a desired confidence interval (Z = 1.96 for 95%CI, Z = 2.58 for 99%CI). After we have calculated the upper and lower limit, we have to find the antilog of these two numbers, by exponenting their values. Similarly, the confidence interval for odds ratio (OR) is calculated as follows:

$$Ln(OR) \pm z_\alpha * sqrt(1/a + 1/b + 1/c + 1/d).$$

As in the previous case, we also have to find the antilog of its lower and upper limit. If a particular confidence interval includes the value of 1, RR or OR indicate that there are no significant differences (or effects).

It is also useful that we report confidence interval for the value of the area under the ROC curve (AUC). The formula is:

$$AUC \pm z_{\alpha/2} * SE(AUC)$$

"SE" stands for the standard error of AUC, and $z_{\alpha/2}$ is a value related to the desired confidence interval. The formula for SE calculation depends on both AUC estimate and sample size. If this interval contains the value of .5, the diagnostic power of the applied test is said to be negligible.

The second measure that is usually omitted by authors of scientific papers is the effect size, usually denoted as *d*. For two groups, the effect size is equal to $M_1 - M_2$, divided by the pooled standard deviation (SD_{pooled}). SD_{pooled} is calculated as the root mean square of the standard deviations of the groups compared [4]. Researchers should also indicate if the effect size is (Cohen, 1988): small (d = .2), medium (d = .5), or large (d = .8).

In ANOVA, the effect size is the ratio of the sum of squares for the effect of interest (SS_{effect}) and the total sum of squares (SS_{tot}). It is called eta-squared coefficient (η^2). If we want to determine partial (specific) effects, we use partial eta-squared coefficient (η_p^2). We obtain it when we divide SS_{effect} by the sum of SS_{error} and SS_{effect}. If $\eta^2 = .01$, it is considered to be small, if $\eta^2 = .06$, it is the medium effect, and if $\eta^2 = .14$ it is a benchmark for large effect (e.g. [11]).

If we consider correlation and regression analysis, r or r^2–coefficient (coefficient of determination) can serve as effect size estimates. In addition, we can interpret R and R^2 (coefficient of multiple correlation as well as of multiple determination, respectively). Cohen [4] proposed that r = .1 indicates small, r = .3 represents medium, and r = .5 indicates large effect size.

Therefore, we have to inform our readers (because some of them are not familiar with the meaning of statistical coefficients) about the strength of correlation/ association between variables or the impact of one variable on another.

Other important topics are related to tabular and graphical representation of results. *The basic principle is that of not displaying the same data using both tables and graphs!* If we break that rule, our paper suffers from redundancy.

Usually, the format/design of graphs and tables obtained in statistical softwares has to be somehow changed (simplified and adjusted). We cannot display tables with lots of decimals, different fonts and a lot of underlining (e.g. emphasizing variables or numbers by using both bold and italic). Furthermore, we have to pay attention to figures, especially graphs. Landau and Everitt [12] recommended that graphical display of results should: have its own title, have titles for different axes, be converted into black and white, not be in a box, and have *y*-axis which starts at the origin. Tables and graphs should be clear and meaningful, containing data needed to structurally inform readers about the obtained results. Table 4 is an example of inadequately summarized results.

Table 4. An example of the data presented improperly

	Birth weight	*Severity of heart disease*	*Obesity in adulthood*
Birth weight	1	**.250**, p < .05	**.420**, p < .001

Coronary heart disease	**.250**, p < .05	1	**.365**, p < .01
Obesity in adulthood	**.420**, p < .001	**.365**, p < .01	1

As can be noticed, variable labelled as "birth weight" is put in both bold and italic. Near the coefficients of correlations, there are the levels of significance. The levels of significance can be shown in tables, but in a separate column, along with other coefficients. In addition, this table contains all data written twice. Moreover, this table contains vertical lines, which can be omitted.

Table 5. An example of the data presented correctly

	Birth weight	Severity of heart disease	Obesity in adulthood
Birth weight	1	.250*	.420***
Coronary heart disease		1	.365**
Obesity in adulthood			1

** p < .05; ** p < .01, *** p < .001*

On the other side, we displayed data correctly (table 5). Every coefficient is marked with one, two, or three asterisks, which indicate its significance at the different levels. This table is simpler and more transparent than was the previous one.

With regard to graphs, lots of researchers mistakenly use bar plots to display arithmetic means. Bar plots (and pie charts) are suitable for presenting frequencies (i.e. the number of participants in subsamples) and percentages that add up to 100%. It is useless to present percentages which do not add up to 100%, because readers could be confused of this. Histograms are usually used to display the distribution of the data (we can see that distribution is approximately normal, multimodal, left-skewed, right-skewed, etc.). Scatter plots are suitable for presenting correlation between two variables visually. We are going to provide an example for improperly (Figure 1) and correctly (Figure 2) presented data. Suppose we want to display frequencies (i.e. the number of participants in different subsamples) of males and females in three groups of participants: those who were diagnosed with obsessive-compulsive disorder (OCD), those whose diagnostic category was posttraumatic stress disorder (PTSD), and participants who were diagnozed with generalized anxiety disorder (GAD).

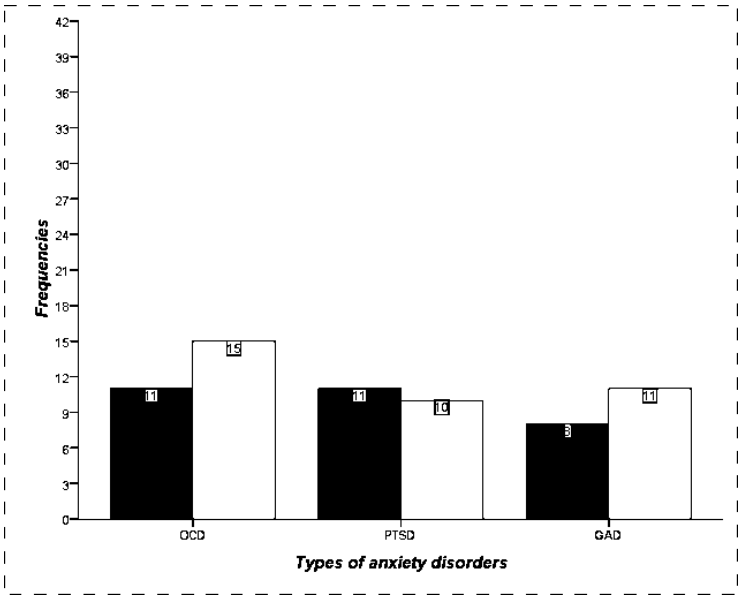

Figure 1. Example of the incorrectly displayed data

As can be noticed (Figure 1), the scale for y-axis includes a very wide range of values (from 0 to 42), which is inappropriate for displaying the obtained frequencies. Additionally, we displayed numerical values in the main frame of our chart and we made the frame of the chart, by using dashed line. The other shortcomings of our chart are the lack of a chart legend as well as over-emphasized titles of *x* and *y*–axis.

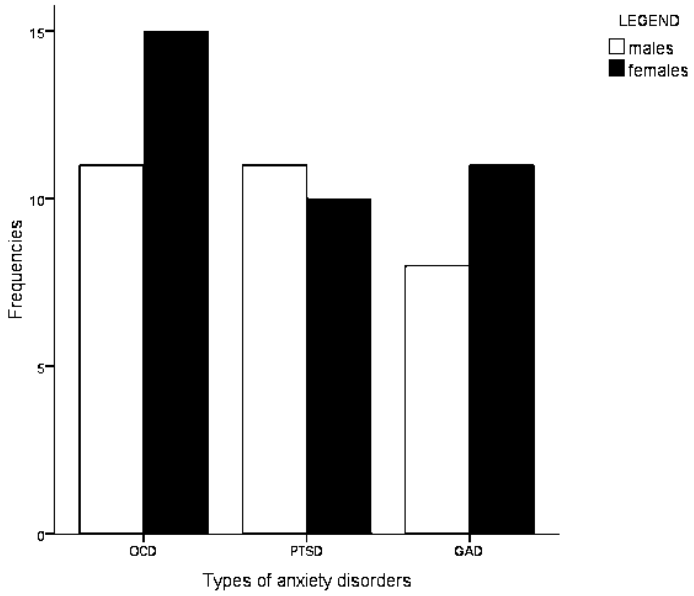

Figure 2. Example of the correctly displayed data

Concerning figure 2, we can see that there is the legend here and the range of values linked to the y-axis is appropriate. The titles of axises are properly formatted and we can clearly read off every value of our interest.

In conclusion, every researcher should report all coefficients, and other values which can help readers to understand a particular scientific article. All values, abbreviations, and symbols have to be explained as well as interpreted properly. Finally, tables and graphs should reflect the obtained results clearly.

Conclusion

Taking into account the previous considerations, the following can be concluded:

1) Values/numbers in the database should be checked for mistakes and the coding of variable values has to be conducted properly.
2) Based on data distributions, every researcher should decide what statistical methods can be applied (parametric or non-parametric).
3) If we decide to use parametric methods, we should exclude outliers or transform our distributions (if they deviate from the bell-shaped curve). On the other hand, if we choose to apply non-parametric methods, we have to state their lack of power for refusing or accepting the null hypothesis.
4) We are also expected to report confidence levels, size effects, degrees of freedom, and briefly explain what these statistical indicators mean.
5) Tables and graphics should contain clearly displayed data that can facilitate readers' orientation in articles.

In addition, researchers in the field of the biomedical sciences (as those in other fields where statistics is crucial for deriving valid and reliable conclusions), should learn the background of every

statistical procedure they use as well as seek for the right ways in which results can be shown to the academic community. Professors of biomedical statistics have to encourage their students to be aware of the importance of statistical education, especially those who plan to engage in conducting quantitative studies.

References

1. A. Agresti and B. Finlay: *Statistical methods for the social sciences* (4th ed.). New York: Pearson, 2009.
2. R. M. Baron and D. A. Kenny: *The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations*. J. Pers. Soc. Psychol. 51(1986), 1173-1182.
3. G. Camilli and K. D. Hopkins: *Testing for association in 2 * 2 contingency tables with very small sample sizes.* Psychol. Bull. 86(1979), 1011-1014.
4. J. Cohen: *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates, 1988.
5. J. G. Eisenhauer: *Degrees of freedom*. Teach Stat. 30(2008), 75-78.
6. J. F. Elder: *Top 10 data mining mistakes: avoid common pitfalls on the path to data mining success.* In: R. Nisbet, J. Elder, and G. Miner (Eds). Handbook of statistical analysis & data mining applications. Burlington, MA: Academic Press, 2009, p. 733-754.
7. M. F. Festing and D. G. Altman: *Guidelines for the design and statistical analysis of experments using laboratory animals*. ILAR J. 43(2002), 244-258.
8. V. Gupta: *SPSS for beginners*. Washington: VJBooks Inc., 1999.
9. J. Han and M. Kamber: *Data mining concepts and techniques*. Morgan San Francisco: Morgan Kaufmann Publishers, 2006.
10. S. K. Kachigan: *Statistical analysis: an interdisciplinary introduction to univariate & multivariate methods,* New York: Radius Press, 1986.
11. D. Lakens: *Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs.* Front. Psychol. 4(2013), p. 863.
12. S. Landau and B. S. Everitt: *Statistical analyses using SPSS*. New York: Chapman & Hall/CRC, 2004.
13. T. Lang: *Twenty statistical errors even YOU can find in biomedical research articles*. Croat Med J., 45(2004), 361-370.
14. D. Moore and G. McCabe: *Introduction to the practice of statistics* (3rd ed.). W. H. Freeman, 1998.
15. J. W. Osborne and E. Waters: *Four assumptions of multiple regression that researchers should always test.* PARE 8(2)(2002), 1-9.
16. M. Pagano and K. Gauvreau: *Principles of biostatistics* (2nd ed.). Belmont, CA: Brooks/Cole, 2000.
17. J. F. Pallant: *SPSS survival manual* (2nd ed.). Sydney: Allen & Unwin, 2005.
18. S. Repišti: *Problem razumijevanja varijance i kovarijance i postupaka njihovog računanja u psihometriji* [The problem of understanding variance, covariance, and the procedures of their calculation in psychometrics], IMO, 4(2012), 31-43.
19. S. Repišti: *ROC krive u psihološkoj statistici na primjeru potencijala osobina ličnosti za distinkciju iznadprosječno i ispodprosječno optimističnih osoba* [ROC curves in psychological statistics in the case of personality traits' potential to distinguish upper- and lower-optimistic people], IMO, 6(2014), 19-30.
20. J. P. Shaffer: *Multiple Hypothesis Testing*. Annu. Rev. Psychol., 46(1995), 561-584.
21. A. M. Strasak, Q. Zaman, K. P. Pfeiffer, G. Göbel, and H. Ulmer: *Statistical errors in medical research – a review of common pitfalls.* Swiss Med Wkly, 137(2007), 44-49.
22. L. Tenjović and S. Smederevac: *Mala reforma u statističkoj analizi podataka u psihologiji: malo p nije dovoljno, potrebna je i veličina efekta* [A small reform in the data analysis in psychology a small p is not enough, effect size is needed too]. Primenjena psihologija, 4(2011), 317-333.
23. N. A. Weiss: *Introductory statistics* (9th ed.). New York: Pearson, 2012.

Received by editors 20.06.2015; Available online 13.07.2015.