COMMON MISTAKES OF STATISTICAL TESTING AND RESULTS INTERPRETATION IN HEALTHCARE STUDIES

Vladimír MELUŠ [1*], Zdenka KRAJČOVIČOVÁ [1], Patrícia BAŇÁROVÁ [1], Miroslav ČERNICKÝ [1], Jana NETRIOVÁ [1,2]

[1] *Faculty of Healthcare, Alexander Dubček University of Trenčin in Trenčín, Študentská 2, 911 01 Trenčín, Slovak Republic*
[2] *St. Michael´s Hospital, Inc., Cesta na Červený most 1, 811 05 Bratislava, Slovak Republic*
*Corresponding author E-mail address: vladimir.melus@tnuni.sk

Received 23. 07. 2013; accepted 16. 08. 2013

Abstract

There are available many statistical software applications nowadays. Their application is many times reduced to filling of data and to the mere choice of statistical test from the menu without an adequate knowledge of the requirements of the selected test. Along with misinterpretation of results it may lead to incorrect conclusions.

The aim of our paper is to demonstrate the most common cases of misuse of statistical tests and misinterpretation of their results. On example of data obtained by measuring the range of shoulder movement we show the importance of correct choice of statistical test according to their requirements of the number of statistical units and distribution of values. We emphasize the proper consideration of outliers with regard of biological relevance of their inclusion in the tested file. The differences between statistical and biomedical interpretation of laboratory analyses we demonstrate on example of NT-proBNP.

We can conclude that despite technical progress there remain risks of statistical analyses primarily in approach and interpretation of testing. Therefore the completion of technical infrastructure for the development of science and research should be related with awareness rising on the basic principles of biostatistical applications.

Keywords: biostatistical tests, interpretation of results, technical infrastructure

1 Introduction

In the past several decades an imposing development of computer technology has been achieved. This has resulted into more rapid and more accurate obtaining of numerical results. Computing technology significantly reduces random errors in calculations. Therefore, sources of errors are seemingly included only in the phase of data collection and interpretation of results.

Experimental design, postulation of zero hypotheses and data collecting are closely linked and interrelated. Any error in one of those steps will also affect others. It often happens that on the basis of the available data people are secondary trying to postulate hypotheses and use statistical tests [1-3].

The level of education of the routine application of statistical methods is currently not satisfactory. Paradoxically, computer processing often leads to vague application of statistical tests without considering their limitations and requirements.

2 Aim

On the example of several small studies we demonstrate the most frequently errors in assessment of normality of distribution, impact of outliers, assessment of correlation and their impact on the interpretation of results.

3 Material and methods

Numerical data were obtained from application research for thesis purposes in biomedicine, especially in physiotherapy and laboratory medicine. Basic statistics of the samples was determined by the arithmetical mean, standard deviation, median, range given by minimum and maximum values and total number of samples. Comparison of two independent samples of data was performed with parametrical t-test, t-test with Welch approximation as well as with non-parametrical Mann-Whitney test [4-7].

We listed all possible results, including examples of bad results and misinterpretations of the numerically correct results. The significance level of the test results was set to $p < 0.05$.

4 Results and discussion

4.1 Outliers and parametrical tests limitations

In physiotherapy, there is often used Rippstein´s plurimeter, a special goniometer, which is able to determinate the upper limb flexion with an accuracy of one degree [8, 9]. It has an important role in physiotherapy of lymphedemas and shoulder injuries, which are after associated with reduced motivity. Nowadays they are new efforts to use physiotherapy in connection with hyperbaric oxygen therapy. In the table 1 we can see the results of such study with relatively small number of patients (n=20). Due to the small number of patients we can see at first sight two crucial facts:

***Table 1** Primary data of patients*

A	6	5	9	8	7	10	90	9	10	11	12	12	11	10	5	8	7	6	9	175
B	17	16	17	18	25	20	24	22	23	21	19	26	25	24	19	22	21	20	23	18

Legend: A – patients, B – control group; total number of patient n=20; data are given in degrees of flexion angle

1/ The patient´s group has apparently lower motility in comparison with the control group. The only question is, whether this difference is statistically significant.
2/ Patient´s set of numbers includes two extremely high values (90 and 175, respectively).

The fundamental question is: Are both outliers intrinsic to the examined fields or not?

There is no problem to solve this issue mathematically with removing both samples. But we must carefully assess, whether the removal of outliers is correct from the nature of the problem. If the answer is: "*not remove*", we will be restricted to use nonparametrical tests only. Parametrical tests (in this case t-test) are limited with the requirements for the normality of values distribution, homogenity of variance (so - called homoskedasticity) and with required number of samples $n > 30$.

***Table 2** Descriptive statistics of tested groups*

Groups	**n**	**X**	**SD**	**Min**	**Max**	**Med**
Patients	20	21.00	40.61	5.00	175.00	9.00
Control group	20	21.00	2.99	16.00	26.00	21.00

Legend: n – number of results, X – aritmetical mean, SD – standard deviation, Min – minimal observed value, Max – maximal observed value, Med – median

Diametrical difference in test results we can see in tables 2 and 3. Arithmetical means of both samples are identical, however this is not their intrinsic property. It is result of both outliers in patient´s sample. The same reason causes extremely high standard deviation of patients, almost twice higher than arithmetical mean (X=21; SD=40.61). The consequence of the misuse of the parametrical t-test and t-test with Welch correction is evident: both tests indicate that the both groups are in their mean values not significant (p> 0.99).

The only correct test is therefore the non-parametrical Mann-Whitney test, which detect real differences of the both groups (p < 0.001). It is notable, that both medians also indicate fundamental differences between groups (Med=9.00 in patients and Med=21.00 in controls) as well as total ranges given by minimal and maximal values (table 2).

***Table 3** Results of statistical tests*

Statistics	Basic statistics		Unpaired t- test		Unpaired t- test/ Welch correction		Mann-Whitney test	
Groups	**X**	**Med**	**d.f.**	**p**	**d.f.**	**p**	**U**	**p**
Patients	21.00	9.00	38	> 0.99	19	> 0.99	40	< **0.001**
Control group	21.00	21.00						

Legend: X – arithmetical mean, Med – median, d.f. – degrees of freedom, p – significance level, U – test characteristics of the Mann-Whitney test

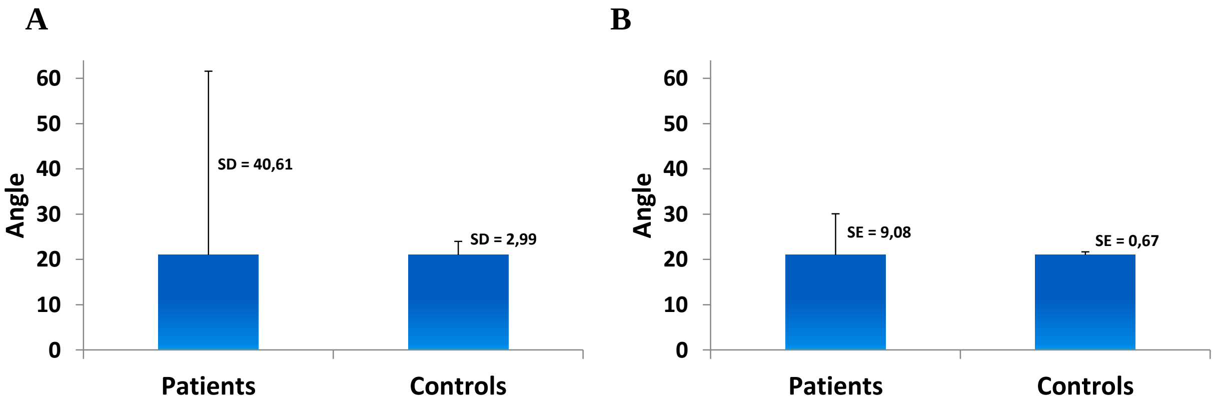

***Fig 1** Differences in graphical depiction of basic statistical parameters:*
***A** Mean and standard deviation of tested groups, **B** Mean and standard error of tested groups*

The notable differences are in the graphical visualization of the results. As we can see on the figures 1A and 1B, both graphs are indeed correct, but subconsciously distort the differences between groups. This is due to the use of standard error instead of standard deviation, which **optically** decreases variability indication and thus evokes the impression of identical samples. Please note, that identical arithmetical means don´t mean identical samples!

In the case of the excluding of both outliers we can use all three statistical tests with the same result (tables 3, 5). Graphical depiction also represents real situation (figures 2A, 2B).

***Table 4** Descriptive statistics of tested groups after outliers removing*

Groups	**n**	**X**	**SD**	**Min**	**Max**	**Med**
Patients	18	8.61	2.25	5.00	12.00	9.00
Control group	20	21.00	2.99	16.00	26.00	21.00

Legend: n – number of results, X – arithmetical mean, SD – standard deviation, Min – minimal observed value, Max – maximal observed value, Med – median

Removing the outliers seems to be very effective for achieving of statistically correct values and their proper interpretation. However, it must also be taken into account from the interpretation perspective relative to nature of the test variable. Outliers may itself be an important trace of an important feature of tested sample group. The risk is that due to their removing we can lose important information.

Table 5 *Results of statistical tests of corrected groups*

Statistics	**Basic statistics**		**Unpaired t-test**		**Unpaired t- test/ Welch correction**		**Mann-Whitney test**	
Groups	**X**	**Med**	**d.f.**	**p**	**d.f.**	**p**	**U**	**p**
Patients	8.61	9.00	36	< **0.001**	34	< **0.001**	0.00	< **0.001**
Control group	21.00	21.00						

Legend: X – arithmetical mean, Med – median, d.f. – degrees of freedom, p – significance level, U – test characteristics of the Mann-Whitney test

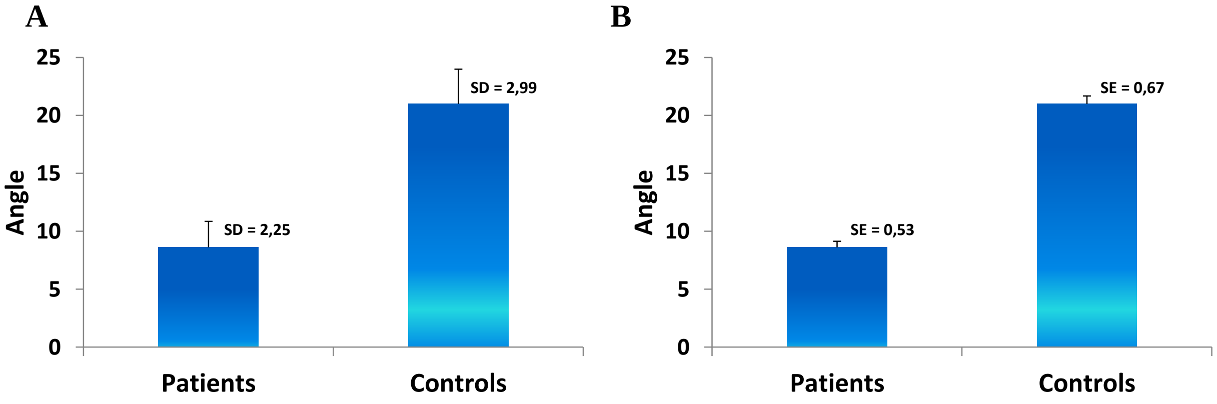

Fig 2 *Differences in graphical depiction of basic statistical parameters after outliers removal: A Mean and standard deviation of tested groups, B Mean and standard error of tested groups*

4.2 Correlation versus conformity of average values

Another important problem consists in the number of samples and the nature of the numerical values. An example would be the determination of NT-proBNP, a significant natriuretic protein, important in the early diagnosis of heart failure. From the results of laboratory measurements of NT-proBNP concentrations listed in table 6 we can see several important facts:

- Although in the study was available only a small number of patients, the used statistical test was chosen correctly. In such a situation, the usage of nonparametric test is a good choice, reducing the probability of misinterpretation.
- Wilcoxon paired test showed statistically significant differences in mean values of both analysers (p=0.03).
- Spearman´s rank correlation coefficient was chosen also properly as a nonparametric alternative to the Pearson´s correlation coefficient.

Despite the above-mentioned facts, which indicate the correct use of statistical tests, there are risks related with the interpretation of these results.

Firstly, we can see, that both analysers are statistically different in the average results, but the difference is only around 2.03%. Are the results of statistical analyzes relevant in view of their relevant biomedical interpretation? From the previous studies we know, that in stable patients after heart failure, the intra-individual variability of NT-proBNP varies in the range of approximately 7% for within-hour intraindividual variability and even around 21 – 47 % for within-week intraindividual variability. With regard of reference limit level (900 pg/ml) we can conclude, that differences between both systems are not so remote for unified assessments. In other words, in this case, we could afford to accept the interchangeability of methods under certain conditions. In both cases the coefficient of variation is relative high (CV=50.45% for E411 and CV=49.09% for Pathfast) and very similar [10-12].

***Table 6** Results of NT-proBNP testing obtained from two different analysers*

Age [years]	**NT-proBNP [pg/ml]***	**Analyser**	**Basic statistics**					**Statistical analyses**	
			n	**X**	**SD**	**Min**	**Max**	**p**	**R**
50 – 75	> 900	Elecsys E411	9	2 774.11	1 399.41	1 041	5 601	**0.03**	1.00
		Pathfast	9	2 831.55	1 390.10	1 075	5 577		

Legend: n- number of samples, X – arithmetical mean, SD – standard deviation, Min – minimal observed value, Max – maximal observed value, p – significance level of nonparametric Wilcoxon paired test, R – Spearman´s rank correlation coefficient, * - reference limit of NT-proBNP concentration.

Another matter is correlation, i.e. the extent of interrelationships between results of both analysers. In this case is correlation coefficient very high (R=1.00). Of course, nonparametric Spearman´s rank correlation coefficient had to be used.

5 Conclusion

Statistical software made available testing to a much wider professional public than it was before the massive deployment of computer technology. This trend is associated with less knowledge of the statistical design, statistical tests as well as the fundamental principles of interpretation of results according to the biological nature of tested parameters. Another common problem is the unsystematic work and effort to statistical re-processing of older data, the number and nature of those do not allow obtaining of reliable results.

Acknowledgement

This publication was cerated in the frame of the project „Dobudovanie technickej infraštruktúry pre rozvoj vedy a výskumu na Trenčianskej univerzite Alexandra Dubčeka prostredníctvom hyperbarickej oxygenoterpie“ ITMS code 26210120019, of the Operational Program Research and Development funded from European Fund of Regional Development

References

[1] G. W. Snedecor, W. G. Cochrane: Statistical methods. Iowa State University Press, Iowa, 1989.

[2] B. Rosner. Fundamentals of Biostatistics, 7th edition. Cengage Learning, Stamford, Connecticut, 2010.

[3] L.M. Sullivan. Essentials of Biostatistics for Public Health, 2 th edition, Jones & Bartlett Learning, Burlington, Massachusetts, 2011

[4] J. Chajdiak: Štatistika jednoducho v Exceli. 1st ed., Statis, Bratislava, 2013.

[5] J. Chajdiak: Štatistika v Exceli 2007. Statis, Bratislava 2009.

[6] J. Chajdiak, E. Rublíková, M. Gudába: Štatistické metódy v praxi. 1st ed., Statis, Bratislava, 1997.

[7] N. Poliaková: Risk factors of allergic diseases, University Review, 2012, Vol. 6, No. 1, p. 66-70.

[8] S. Green, R. Buchbinder, A. Forbes, N. Bellamy: A standardized protocol for measurement of range of movement of the shoulder using the Plurimeter-V-inclinometer and assessment of its intrarater and interrater reliability, Arthritis & Rheumatism, 1998, Vol. 11, No. 1, p. 43-52.

[9] L. Watson, S.M. Balster, C. Finch, R. Dalziel. Measurement of scapula upward rotation: a reliable clinical procedure. British Journal of Sports Medicine, Vol. 39, No. 9, p. 599-603

[10] R. O´Hanlon, P. O´Shea, M. Ledwidge, C. O'Loughlin, S. Lange, C. Conlon, D. Phelan, S. Cunningham, K. McDonald: The biologic variability of B-type natriuretic peptide and N-terminal pro-B-type natriuretic peptide in stable hearth failure patients, Journal of Cardiac Failure, 2007, Vol. 13, No. 1, p. 50-55.

[11] A. H. Wu: Serial testing of B-type natriuretic peptide and NT-proBNP for monitoring therapy of heart failure: The role of biologic variation in the interpretation of results, American Heart Journal, 2006, Vol. 152, No. 6, p. 828-834.

[12] H. Reinhard, P.R. Hansen, N. Wiinberg, A. Kjær, C.L. Petersen, K. Winther, H. H. Parving, P. Rossing, P. K. Jacobsen: NT-proBNP, echocardiographic abnormalities and subclinical coronary artery disease in high risk type 2 diabetic patients. Cardiovascular Diabetology, 2012, vol 11:19, p 1-10.

Review: Peter Božek
Jana Slobodníková