Confidence and tolerance intervals – a tool for biomedical data analysis aimed at clear evidence

MIROSLAV MIKULECKY
Bratislava, Slovak republic

MIKULECKY M. **Confidence and tolerance intervals – a tool for biomedical data analysis aimed at clear evidence.** Cardiol 2004;13(4):211–215

In domestic as well as foreign publications standard deviation or standard error is often used to express statistical uncertainty, sometimes even as an undefined number. These indices possess a probabilistic interpretation only indirectly, after respecting the number of measurements. That is why they have to be substituted e.g. by intervals of 95% confidence for mean and by those of 95% tolerance for the individual. As the main tool of evidence the *p* value is usually used. Its information value, however, represents a mutually indistinguishable mixture of the effect size and its precision. Besides, it leads to the false impression of a lower risk than there really is. The best characterization of the size of the effect and of its precision is the confidence (for mean) or tolerance (for individual) interval of the difference as compared with the null value on one side, and with the minimal acceptable difference on the other. Using the *p* value, the replication probability – i.e. saying that a repetition of an observation or experiment will yield the same result – can be calculated. This probability makes the interpretation of *p* value appropriate. Adhering to old, no longer suitable schemes of statistical description can be explained by the lack in the education in biometrics at pre- and postgraduate level of medical teaching. Orientation towards inferential, conjectural statistics and arriving at conclusions together with their probability interpretation is needed. This knowledge is recently considered to be one of literacy – the ability to "read" the sense of numbers. A physician-researcher and physician in practice will, for example, more effectively communicate with a drug producer, to the advantage of all four participants of the process, including the patient.

Key words: Evidence – Confidence – Tolerance – *p* value – Replication

MIKULECKÝ M. **Intervaly spoľahlivosti a tolereancie – hlavný nástroj dôkazu pre analyzovanie biomedicínskych údajov.** Cardiol 2004;13(4):211–215

V domácich aj zahraničných publikáciách sa často k vyjadreniu štatistickej neistoty používa smerodajná odchýlka alebo stredná chyba, niekedy dokonca nedefinované číslo. Tieto ukazovatele majú pravdepodobnostnú interpretáciu iba nepriamo, pri rešpektovaní počtu meraní. Preto ich treba nahradiť intervalmi, napríklad 95 % spoľahlivosti (konfidencie) pre priemer a 95 % tolerancie pre jednotlivca. Ako hlavný nástroj dôkazu sa väčšinou používa *p* hodnota. Jej informačná hodnota však predstavuje vzájomne nerozlíšiteľnú zmes veľkosti efektu a presnosti jeho stanovenia. Okrem toho vytvára klamný dojem menšieho rizika aké v skutočnosti je. Veľkosť efektu a jeho presnosť najlepšie vyjadruje konfidenčný (pre priemer) alebo tolerančný (pre jednotlivca) interval pre rozdiel v porovnaní jednak s nulovou hodnotou, jednak s minimálnym prijateľným klinickým efektom. Z *p* hodnoty možno vypočítať replikačnú pravdepodobnosť – že totiž pri opakovaní pozorovania alebo pokusu dostaneme rovnaký výsledok. Táto pravdepodobnosť uvádza interpretáciu *p* hodnoty na pravú mieru. Zotrvávanie na starých, už dávno nevyhovujúcich šablónach štatistickej deskripcie možno vysvetliť chýbajúcim vzdelávaním v biometrike na pre- i postgraduálnej úrovni výučby medicíny. To treba orientovať na štatistiku inferenčnú, úsudkovú, ktorá formuluje závery spolu s ich pravdepodobnostnou interpretáciou. Jej znalosť sa dnes pokladá za jednu z gramotností – schopnosť „čítať" zmysel čísel. Takto pripravení lekár-výskumník a lekár-praktik budú napríklad účinnejšie spolupracovať s výrobcom lieku na prospech všetkých štyroch účastníkov procesu, vrátane pacienta.

Kľúčové slová: dôkaz – spoľahlivosť – tolerancia – *p* hodnota – replikácia

Recently, I was asked to write an Editorial concerning the present status of dealing with the scientific heritage of the Academician Ladislav Dérer (1). I stressed that its substantial component is an appropriate, updated statistics. I expressed concerns as to the fulfilling of this approach.

The issue is real, not only in the medical literature of this country but worldwide, except for a few "torchbearers" such as the British Medical Journal (BMJ). In fact, often only descriptive, not inferential (mathematical, analytic, inductive, evaluating) statistics, which tell unequivocally of the real probabilities of each conclusion, valid for corresponding *population,* is used. Many papers persevere with the *description of samples* with the aid of means and standard deviations, standard errors, or even some numbers which are *not* specified! The usually added *p* value represents no practical probability – it is bound on the hypothetical validity of the null hypothesis, neglecting the infinite number of possible alternative hypotheses. Moreover, the probability of obtaining an effect in an individual – the core of clinical medicine – is totally omitted.

Those using descriptive statistics of samples are at the intellectual level of the end of 19th century. In those times, W. S. Gosset (known later as "Student") and R. A. Fisher recom-

From 1st Medical Clinic, Teaching Hospital, Comenius University, Mickiewiczova 13, 813 69 Bratislava, Slovak republic
Manuscript received December 12, 2003; accepted for publication March 31, 2004
Address for correspondence: Prof. Miroslav Mikulecký, I. interná klinika, Mickiewiczova 13, 813 69 Bratislava, Slovakia, e-mail: biometrik@pobox.sk

mended "dealing with real data inductively, knowing that practical action will be taken on the basis of your conclusion" (2).

In the following critical contribution, the problem will be simplified on univariate investigation of a drug effect to show the practical advantages of the proposed course of the clearly and practically formulated output.

Example

A physician, reading a paper about a drug effect, should be interested in the following four meaningful features.

1. What *average* effect can be expected? It should be given as percentage change, e.g. as the decrease of cholesterolemia by –25%, i.e. from the starting value of 100% to 75 %.

2. The point estimate sub 1 alone, however, has zero probability. We have therefore to ask what range of this average effect can be predicted from the n-sized sample for the underlying population with a chosen reasonable probability, e.g. 95%? Such interval is called *confidence* interval and could extend in the given example from –24% to –26% decrease.

3. In clinical medicine, however, the main interest is oriented towards an individual, not towards the average. The corresponding interval is called the *tolerance* interval and will, of course, be much wider. In the given example, the 95% tolerance interval with 95% confidence could be between –59% decrease and even +10% increase.

4. Such conclusions can be drawn from one sample. There arises then the question about the probability that repeated trial will give the same result. This *replication* probability is hidden in the information sub 1 – 3. It gives, however, a clearly explicit idea about the reliability of these statements.

A practical physician, informed in this inferentially statistical manner, will be able to evaluate exactly his own experience with the drug. Thus, he will be the qualified partner of corresponding investigators and of the drug producer, in *favour of all four participants in the drug business,* including the patient.

Unfortunately, this model appears nowadays as a "fata morgana". In reality, there are often given e.g. both starting and final means (x) ± standard deviations (*SD*), standard errors (*SE*) or even an undefined number, together with the *p* value for significance. With sample sizes *n* such information (*n, x, SD* or *SE, p*) is complete. Nevertheless, it is not optimal: the uncertainty of the difference, i.e. of the effect, is not evident, the difference cannot be compared appropriately with another difference, and the *p* value hides several pitfalls. At other times, particularly in short information texts about a drug, only the average decrease *d* after therapy is given, usually but not always with the *p* value. Such a concise description with the *p* value is, together with sample sizes (*n, d, p)*, also complete, and seemingly appears very clear. The problems with it, however, are the same as those with the preceding form of output. The reader of such information will hardly have a sufficiently formed idea about the effect he can really see in his patients. Errors of interpretation are threatening here, "with misunderstandings of the meaning of the P-values being especially common" (3). "The *p* value is often criticized on the ground that clinicians have difficulty interpreting it or are likely to misinterpret it" (4). Our recent attempt to ascertain the level of corresponding knowledge in Slovak physicians (5) failed: nobody answered the published questionnaire about the *p* value – the ubiquitously used tool of "evidence".

In fact, the *p* value itself is a quite inappropriate measure of evidence. It does not answer the question "on average, how great is the change produced by the intervention?" nor the question "with what precision has the average change been estimated?" (6). According to the latter authors, these questions "are answered by the calculation of confidence intervals, whereas hypothesis testing (*p* values, comment by M.M.) can give only the answer "yes" or "no" to the question "Is there a change?". Accordingly, "the *p* value, a cornerstone of traditional statistics, is inadequate as an evidential measure" (7). Advocates of the *p* value in meta-analysis (8) object that "most practicing statisticians consider the *p* value to be a useful inferential measure". In a purely theoretical controversial debate, Goodman (9) defends his view. Morgan also has a negative attitude towards *p* values (10): "…with the efflorescence of computers and proliferation of databases *p* values threatened to overwhelm the data". He speaks about "a ruthless search for significance, whose success (almost always based on the magic value of $p < 0.05$) would 'validate' the hypothesis...". In contemporary England, where such torchbearers of classical biometry as Pearson Sen, Student and Fisher worked, "The BMJ (British Medical Journal, comment by MM) now expects scientific papers submitted to it to contain confidence intervals when appropriate" (11). It also "wants a reduced emphasis on the presentation of P values from hypothesis testing" (12). The rationale for using confidence intervals is, according to Gardner and Altman (13), that "Presenting study findings directly on the scale of original measurement, together with information on the inherent imprecision due to sampling variability, has distinct advantages over just giving P values usually dichotomised into *significant* or *non-significant*". Similar policy was announced by Bulpitt (14) in the Lancet. He also mentions confidence limits for the difference between two results. The issue has interesting historical roots (15): the giant of statistics – "Fisher was skeptical about the use of statistics for hypothesis testing (i.e. *p* values, comment by M.M.); he favoured the use of sample statistics for estimation of population parameters".

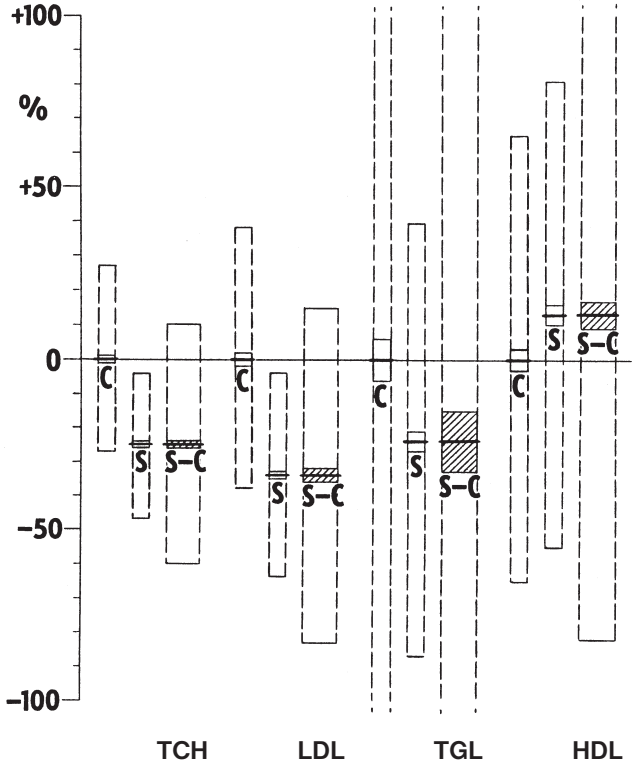

Figure 1 Control (C) and post-SIMVACARD^R (S) values with their differences (S-C) for cholesterolemia (TCH), LDL-cholesterolemia (LDL), triglyceridemia (TGL) and HDL-cholesterolemia (HDL), given as percentage deviation from the control level. Means, shown as the short heavy horizontal abscissae, are accompanied by symmetrical 95% confidence (narrower, for the differences shadowed) and 95% tolerance with 95% confidence (wider) intervals.

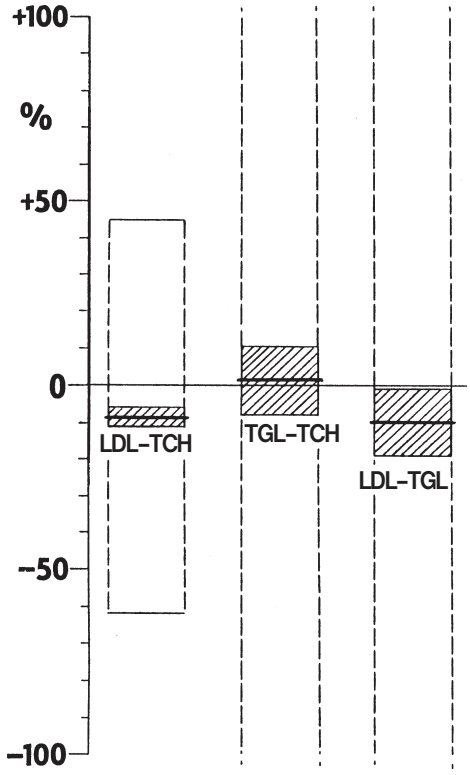

Figure 2 Analogy of Figure 1 for the differences between separate effects

The criticism of the *p* value is oriented against its use as a major explicit indicator, not against its implicit importance after recalculation. Let us demonstrate this on the postregistration follow-up of SIMVACARD^R (16). For that, the sign ±, not specified by the original author, has to be decoded: it means obviously the standard deviation. We have to consider the *unpaired c*-test (the *t*-test is usually used for sample sizes not exceeding 200; *c* is the random variable of the standardized normal distribution) because characteristics needed for the paired *c*-test (mean differences and their standard deviations) were not given. The *p* values are given as their upper limits. Exact *p* values will therefore be calculated.

The cited author gives correct average percentage differences. Their confidence intervals will be computed from the given initial and final mean ± standard error (equal standard deviation divided by squareroot of the sample size, i.e. of 404). The *standard error of the difference* is then computed as the squareroot from the sum of both squared standard errors – before the treatment and after it. The half-width of the 95% confidence interval equals this standard error of the difference multiplied by the *c* or *t* value, found in the tables for the probability of 5% (100% – 95%) and corresponding degrees of freedom (sum of sample sizes of both samples minus 2). The probability is usually given in tables as 2P = 0.05 (17). In our example, the value $c = 1.96$ is appropriate. The resulting confidence interval will be defined as the mean difference ± its half width. The confidence interval for initial and final mean values will be calculated analogically – simply by multiplying their standard error by the mentioned *c* value.

The resulting point and interval estimates for the four biochemical parameters tested, after transformation to percentage deviation from the starting value, are shown in **Figure 1** as starting (control) 0% change (in fact, it corresponds with 100% starting level), as final (after SIMVACARD^R) % change and their difference. Note that despite the same point estimate of % value in the last two cases, the confidence intervals are more or less different. A significant change on the level $\alpha = 0.05$ is shown as non-overlapping of the zero value by the confidence interval for difference.

Besides confidence intervals, 95% tolerance intervals can also be estimated. The procedure is analogical to that for confidence, except for using standard deviations instead of

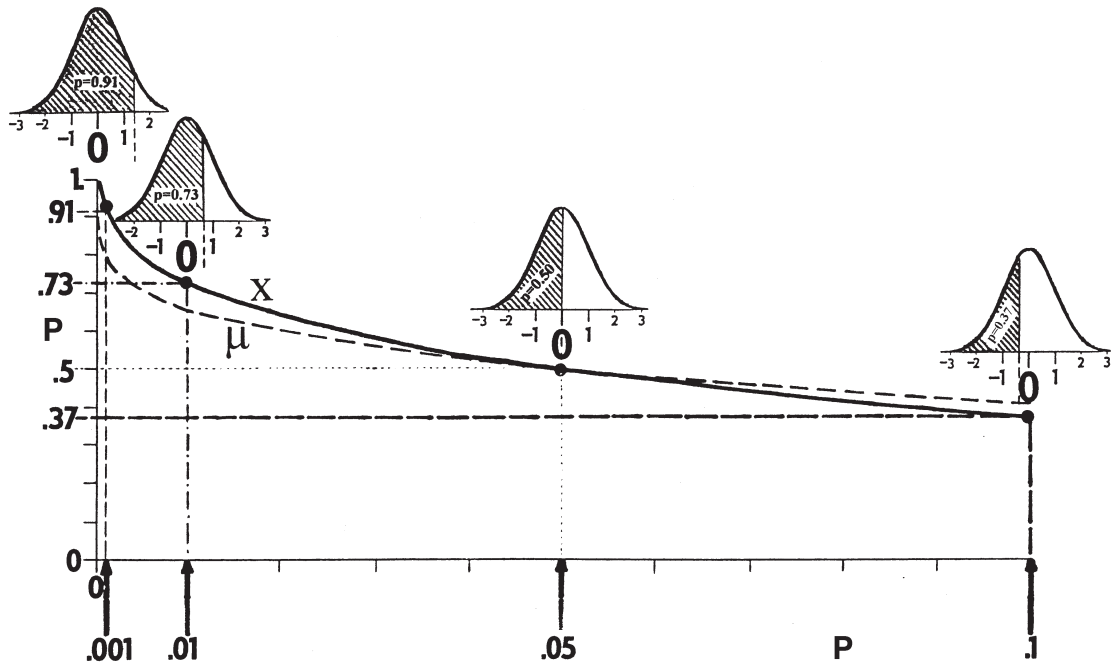

Figure 3 Replication probability p (ordinate), calculated for $\alpha = 0.05$ and plotted versus the P value from the first test (abscissa), as obtained by the simple (x; shadowed areas under four Gaussian curves, corresponding to the replication probability, are shown in four cases) and more exact (μ) procedure (5)

standard errors and, for tolerance with confidence, the tolerance factor k_7 (17) instead of *c* or *t* value. The resulting tolerance limits are substantially – sometimes absurdly – wider than the confidence ones, allowing large excursions for an individual patient (**Figure 1**).

Figure 1 shows that favourable significant changes occurred after SIMVACARD^R treatment in each of the four parameters, with maximum extent in LDL cholesterol. These changes, however, are not quite unequivocal in individual patients.

The confidence and tolerance approach stimulates the posing of further questions. For example, are there significant differences between the drug-lowering effects on the separate four studied biochemical parameters? **Figure 2** gives the answer. The lowering effect on LDL cholesterol is significantly better than that on cholesterol as well as on triglycerides. Nevertheless, the opposite can be true in some individuals.

Scientific papers usually work with statements concerning means, e.g. with the aid of *p* values or confidence intervals. The latter are therefore narrow, particularly for large sample sizes. No wonder, then, that many research results are mutually contradictory (18). If tolerance intervals were compared, the frequency of contradictory conclusions should decrease.

For proper understanding of the confidence and tolerance intervals, exact definitions (17) are given as follows. All are based on the premise that very (infinitely) many samples of the same size are taken from the same stable population. Then, the two-sided 95% confidence interval will enclose the true value of the unknown parameter (e.g. mean) on the average in at least 95% of cases. The sample tolerance interval without confidence probability will enclose on the average 95% of the population, while that with 95% confidence probability will include at least 95% of the population in an average of 95% of cases.

From the data of the discussed paper (16), the exact *p* values were obtained by dividing the average difference by its standard error, so arriving at the values of the random variable *c*. The latter serves for finding out the *p* value from tables (17) or, as in the present case, with the aid of a special statistical pocket calculator (Texas Instruments programmable 58, Solid State Software). The following values of *c* and *p* were identified: $c = 29.6747$ ($p <<< 10^{-9}$) for cholesterol, $c = 29.4136$ ($p <<< 10^{-9}$) for LDL cholesterol, $c = 6.9126$ ($p < 10^{-9}$) for triglycerides and $c = 5.8334$ ($p = 6.6 \cdot 10^{-9}$) for HDL cholesterol. We have now more exactly structured *p* values than those given by the original author as uniform value of $p < 0.0001$. They can be used for estimating the replication probability (19), defined as the probability that repetition of the same experiment or observation will bring result in the same direction. Its value is calculated from the *p* value in two manners (**Figure 3**). It can be surprising – even the author of the procedure was surprised – that from the *p* value of 0.05 a

replication probability of mere 0.5 is derived. For $\alpha = 0.05$ and $p = 0.001$, the simple procedure will arrive at the replication probability of around 0.91 while the more exact method will arrive at approximately 0.8. Accordingly, the impression of an extremely small risk with $p = 0.001$ is false. According to Goodman (19), the *p* values obviously "overstate the evidence against the null hypothesis". In our example with antihyperlipidemic treatment, the extremely low *p* values mean that the replication probability approaches 1. The mean effects appear therefore as quite unequivocally favourable.

For a meaningful medical evaluation of an effect, the minimal relevant change – the "smallest clinically worthwhile difference" (20) – has to be defined. Unfortunately, this is rarely done. It will be, for example, surely not 1 mmHg in the case of antihypertensive therapy, but perhaps already a depression by 5 mmHg would be acceptable. Statistical significance has therefore to be completed by medical significance. According to Morgan (10), confidence intervals "help to change the focus from statistical significance to clinical significance". Corresponding decisions will be made easily using the position of the obtained confidence interval for difference towards the zero difference and towards the minimal relevant change. If, for example, the confidence interval for difference will be situated between the zero and minimal relevant effect lines, without overlapping them, the result will be, despite statistical significant difference, clinically nonsignificant (even statistically significantly!), and a repetition of the trial would not bring much hope of obtaining clinical significance.

It is hardly understandable why many domestic as well as international journals allow the expression of statistical uncertainty with the aid of ± standard deviation or standard error attached to the average value. Moreover, it is not always specified into standard deviation or standard error. According to Altman and Bland (3), "non-reporting of this information has been found in as many as 19% of papers... Some journals, including the British Medical Journal and the Lancet, do not allow this unhelpful usage...". Recently, from 78 abstracts, presented at the VIIIth Congress of the Slovak Cardiologic Society, October 2003 (21), 18 of them were found to have 131 undefined ± signs. Only in one abstract was the standard error and in two abstracts was the standard deviation specified. The latter was in one case peculiarly combined with the median. As a tool of evidence the *p* value served 86-times in 24 abstracts.

It is obvious that the inferential, mathematical statistics, arriving systematically at conclusions with defined probability, should be taught in medicine at both undergraduate and postgraduate levels in Slovakia also. This knowledge has been recently called "numerical literacy" – "the ability to follow and understand numerical arguments" which "is important to everyone" (22).

References

1. Mikulecký M. Editorial. Lek Obz 2004;53:39.
2. Fisher BJ. Guinness, Gosset, Fisher, and small samples. Stat Sci 1987;2:45–52.
3. Altman DG, Bland JM. Improving doctors□ understanding of statistics. J R Statist Soc A 1991;154: 223–267.
4. Diamond GA, Forrester JS. Clinical trials and statistical verdicts: probable grounds for appeal. Ann Intern Med 1983;98:385–394.
5. Mikulecký M. Statistical significance versus replication probability: an enlightening dilemma with clinically oriented solution (and with quiz for those who are interested). (In Slovak). EuroRehab 2002;12:224–230.
6. Evans SJW, Mills P, Dawson Jane. The end of the p value? Br Heart J 1988;60:177–180.
7. Goodman SN. Meta-analysis and evidence. Control Clin Trials 1989a;10:188–204.
8. Zucker D, Yusuf S. The likelihood ratio versus the *p* value in meta-analysis: where is the evidence? Control Clin Trials 1989;10:205–208.
9. Goodman SN. Response to commentary of Zucker and Yusuf. Control Clin Trials 1989b;10:209–210.
10. Morgan PP. Confidence intervals: from statistical significance to clinical significance. Can Med Assoc J 1989;141:881–883.
11. Langman MJS. Towards estimation and confidence intervals. Br Med J 1986;292:716.
12. Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. Br Med J 1986;292:746–750.
13. Gardner MJ, Altman DG. Estimating with confidence. Br Med J 1988;296:1210–1211.
14. Bulpitt JC. Confidence intervals. The Lancet 1987;i:February 28:494–497.
15. Brown GW. Errors, type I and II. Am J Dis Child 1983;137:586–591.
16. Tkáč I. Post-registration follow-up of the preparation SIMVACARD^{R} during secondary prevention.(In Slovak.) Cardiol 2002;11:K/C99.
17. Diem K, Seldrup J.(C. Lentner, ed.) Geigy scientific tables. Vol. 2. Introduction to statistics. Statistical tables. Mathematical formulae. 8th ed. Basle:Ciba Geigy 1982:240.
18. Goncalvesová E. Are we prepared for medicine based on evidence? (In Slovak.) Cardiol 2003;12:161–162.
19. Goodman SN. A comment on replication, p-values and evidence. Stat Med 1992;11:875–879.
20. Daly LE. Confidence intervals and sample sizes: don□ t throw out all your old sample size tables. Br Med J 1991;302:333–336.
21. VIIIth Congress of Slovak Cardiology Society 9. – 11. October 2003. (Abstracts). (In Slovak) Cardiol 2003;12(Suppl. 1):9S–25S.
22. Moore DS, McCabe GP. Introduction to the practice of statistics. New York, Oxford: W. H. Freeman and Co. 1989:790.