STATISTICAL METHODS

Statistical hypothesis testing — how exact are exact p-values?

Gasko R

*Vzajomna zdravotna poistovna Dovera, Kosice, Slovakia.***gasko@mediclub.sk**

Abstract

***Objectives and background:* When testing a hypothesis statistically, a principle is generally accepted that exact p values shall be stated in the treatise. Researchers have the choice of many statistical computer programmes with implemented hypothesis tests. Are exact p values calculated in the same statistical tests by diverse statistical programmes identical?**
***Methods:* The respective zero hypothesis were tested in 5 artificially created data sets by the parametric unpaired t-test, non-parametric Mann-Whitney test, two-tailed F-test. The calculations were carried out by the following programmes: Statistix, version 7.1 (source www.statistix.com), Analyse-it, version 1.62 (source www.analyse-it.com), MedCalc, version 6.14 (source www.medcalc.be). The p values in the same tests were mutually compared.**
***Results:* All three programmes calculated identical exact p values for the t-test. In the remaining two tests in case of 26 out of 44 calculations (59.1 per cent; 95 per cent confidence interval 43–73 per cent) different p values were calculated. The greatest difference was 18.35 per cent. In two cases the values oscillated about 0.05 and this fact caused essentially different interpretation of results.**
***Conclusions:* Using the significance test in the biomedical research has been subject to criticism for a longer period of time. The testing of the zero hypothesis on the arbitrary significance level of 0.05 should be substituted by other methods. Our discoveries should undermine the ungrounded belief of the users of statistical tests – physicians in ununderminable accuracy of mathematical procedures. The use of confidence intervals deems much more suitable although there are objections against them as well.**
(Tab. 4, Fig. 1, Ref. 19.)
Key words: statistics, testing of hypotheses, exact p-values, comparison of statistical programmmes, confidence interval.

The testing of hypotheses by means of suitable tests (Greenhalgh, 1997) serves an answer to the question as to whether the difference between two or more basic sets represented by given selections really exists or whether there is a relation between them (Mikulecký, 1993). The measure of evidence for testing is the p probability value. As a limiting value – the significance level for accepting or refusing the zero hypothesis – of the basic tested hypothesis – the values of 5 % and 1 % are usually chosen. Why the significance values of 5 % and 1 % and not let us say 3 %? Before the era of computers there were statistical charts available for calculations where – regarding the difficulty of calculation – critical values of testing statistics only for some levels of significance, most frequently for levels of 5 % and 1 % were stated. It is easy to remember that in case of normal distribution it is valid that approximately the probability of 5 % corresponds with the distance of two standard deviations. The fact that even the significance levels of 5 % ($p<0.05$) and 1 % ($p<0.01$) are used, is given by the prevailing habit from the past and not by a rule supported by the statistical theory.

Statistical computer programmes are able to calculate exact p values for any statistical tests. In treatises the exact p values should be stated (Mikulecký, 1993; Haas, 1998; Sterne, 2001). If it is not the case, relevant information is lost. The result of $p=0.0003$ is different from $p=0.048$. In the majority of cases the

Vzájomná zdravotná poisťovňa Dôvera, pobočka Košice

Address for correspodence: R. Gasko, MD, VZP Dovera, Murgasova 3, SK-043 25 Kosice, Slovakia.
Phone/Fax: +421.55.6221349

Dedication: The article is dedicated to Prof. M. Mikulecky, MD, DSc.

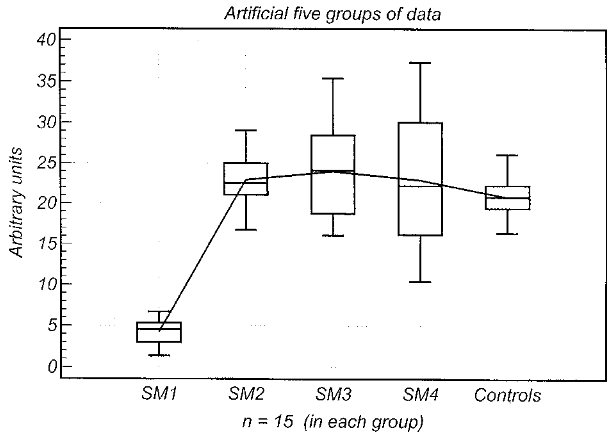

Fig. 1. Expression of intracellular cytokines IFN-gamma in 5 groups (SM1—SM4 and Controls), evaluated as percentage of all leucocytes, box and whiskers plots. Medians are connected with the curve.

works still do not keep to this principle even in the renowned Slovak magazines (example Lazurova et al, 2001).

Exact p values keep up to the appearance of mathematical exactness. The market offers many specialised statistical programmes. Are p values of identical statistical tests calculated by means of different programmes identical?

Methods

As source data values of expression of gamma IFN intracellular cytokines in T cells were used, measured in 4 groups of patients with multiple sclerosis, different in form of illness and treatment – groups SM1 to SM4, and in the control group with different neurological illnesses – Control group (Klímová et al, 2003; Klímová, 2002). The numbers of members of the individual groups were unified for the needs of this study to n=15 and the values were slightly modified to meet the conditions of normal distribution. These are the artificially created sets. The normality of distribution in each group was verified by Shapiro—Wilk, Anderson–Darling and Kolmogorov–Smirnov tests (programme Analyse-it) (Fig. 1).

Tab. 1. Statistical analysis was performed by the unpaired t-test with three statistical programs, and p-values are given. Each program uses different terminology.

Statistix: Two Sample t test Equal Variance

	SM1	SM2	SM3	SM4
SM1	-	-	-	-
SM2	0.0000	-	-	-
SM3	0.0000	0.5400	-	-
SM4	0.0000	0.9942	0.6798	-
Controls	0.0000	0.0559	0.0596	0.3477

Analyse-it: Independence samples t-test, 2-tailed p
MedCalc: t-test, unpaired

	SM1	SM2	SM3	SM4
SM1	-	-	-	-
SM2	<0.0001	-	-	-
SM3	<0.0001	0.5400	-	-
SM4	<0.0001	0.9942	0.6798	-
Controls	<0.0001	0.0559	0.0596	0.3477

The comparison of each group with each zero hypothesis was tested for the difference in means (medians, variance) between two groups by the following statistical tests: comparison of arithmetical means by the parametric unpaired t-test, comparison of medians by the nonparametric Mann–Whitney test, comparison of analysis of variance by two-tailed F-test.

The calculations of statistical tests were carried out by means of the following software: Programmme Statistix, version 1. Student edition, and version 7.1, 1996, 2001 Analytical software, Tallahassee, FL, USA (free trial version, source www.statistix.com). Programmme Analyse-it, version 1.62, © 1997–2001 Analyse-it Software, Ltd., Leeds, UK (free trial version, source www.analyse-it.com). Programmme MedCalc, version 6.14, © 2002 Frank Schoonjans, Belgicko (free demo version, source www.medcalc.be).

Results

When selecting the tests, we met with a serious terminological ambiguity in the naming of the tests in the individual programmes (Tabs 1–3). The names of tests at graphs are taken from the used programmes. MedCalc does not calculate exact values in the Mann–Whitney test. In other 2 tests it calculates the values with 3 decimal positions. Analyse-it and Statistix calculate all 3 tests with 4 decimal positions. Statistix gives the results of p=0.0000 at the value of p<0.0001, which is incorrect and cause incorrect interpretation on the “absolute confirmation” of the zero hypothesis.

In two cases the calculated p values oscillated about the value of 0.05. The use of diverse programmes can – when interpreted rigidly – cause different conclusions as to the statistical significance of differences between the tested sets (Tab. 4).

Where exact p values were calculated, in 26 cases out of 44 (59.1 %; 95 % confidence interval 43–74 %), the results were different, with the difference from 0.0001 to 0.1835. The greatest differences between p values in the same tests measured by diverse programmes are 0.1835, which is 18.35 %.

Discussion

The basic principle of the usage of statistical hypothesis tests is to select a test or tests correctly. The selection criterion is the formulation of the zero hypothesis, less frequently that of alternative hypotheses, the type of data distribution in the tested set according to which parametric or nonparametric tests shall be used, and the number of set members. In our artificially created sets with normal distribution parametric tests, as well as nonparametric tests were used. Only the parametric t-test as has brought identical results of p values in all programmes. In case of other two tests, in cases when exact p values were calculated,

Tab. 2. P-values for Mann-Whitney test. Values oscillating about 0.05 are bold.

Statistix: Rank Sum Two-Sample (Mann-Whitney) Test, Two-tailed p-value for normal approximation

	SM1	SM2	SM3	SM4
SM1	-	-	-	-
SM2	0.0000	-	-	-
SM3	0.0000	0.6482	-	-
SM4	0.0000	0.7716	0.6482	-
Controls	0.0000	**0.0421**	0.1466	0.4429

Analyse-it: Mann-Whitney test

	SM1	SM2	SM3	SM4
SM1	-	-	-	-
SM2	<0.0001	-	-	-
SM3	<0.0001	0.6529	-	-
SM4	<0.0001	0.7748	0.6529	-
Controls	<0.0001	**0.0502**	0.1485	0.461

MedCalc: Mann-Whitney test

	SM1	SM2	SM3	SM4
SM1	-	-	-	-
SM2	<0.0001	-	-	-
SM3	<0.0001	>0.10	-	-
SM4	<0.0001	>0.10	>0.10	-
Controls	<0.0001	**<0.10**	>0.10	>0.10

Tab. 3. P-values for F-test. Values oscillating about 0.05 are bold.

Statistix: One-Way analysis of variance, equal variances

	SM1	SM2	SM3	SM4
SM1	-	-	-	-
SM2	0.0073	-	-	-
SM3	0.0000	0.0422	-	-
SM4	0.0000	0.0015	0.2017	-
Controls	**0.0679**	0.3494	0.0040	0.0001

Analyse-it: Parametric F test, 2-tailed p

	SM1	SM2	SM3	SM4
SM1	-	-	-	-
SM2	0.0073	-	-	-
SM3	<0.0001	0.0422	-	-
SM4	<0.0001	0.0015	0.2017	-
Controls	**0.0679**	0.3495	0.0040	<0.0001

MedCalc: Variance ratio test (F test)

	SM1	SM2	SM3	SM4
SM1	-	-	-	-
SM2	0.003	-	-	-
SM3	<0.001	0.018	-	-
SM4	<0.001	<0.001	0.093	-
Controls	**0.029**	0.166	0.001	<0.001

there were differences, ranging from major to fundamental differences. Using the Bonferroni adjustment (Bland and Altman, 1995; Perneger, 1998), where the p values calculated from several statistical tests are multiplied and put into relation, possible differences in the measured p values are multiplied. Due to the above reasons, we are in agreement with the strict recommendation (Sterne and Smith, 2001) for the purpose of decreasing the risk of incorrect decisions of authors of treatises as well as readers not to use the description "statistically significant, nonsignificant" in their treatises but only the exact p value. The lower the p value the stronger the evidence of the zero hypothesis validity.

We did not have the opportunity to compare according to algorithms that the tests in the programmes are calulated. That is why we cannot justify the differences discovered in the results of the p-calculation. (Hypothetically, it can be for instance an exchange of one-tailed or two-tailed F-tests when developing and describing the programme.) We do not consider it essential for the purposes of this study. The researcher selects the test from the programme according to its name.

The relation between the statistical testing of the hypothesis and the confidence intervals in the clinical research as well as in the laboratory and experimental research has been discussed for many years (Henderson, 1993; Cleofas and Zinderman, 2001; Grimes and Schulz, 2002). The contribution of the significance tests is doubted (Chia, 1997; Evans et al, 1988; Goodman, 1999), the confidence intervals are generally preferred (Gardner and Altman, 1986), although there are objections against them as well (Feinstein, 1998). Confidence intervals provide more information. If there is a 95 % confidence interval for the average glycemia value in the set of patients (5.1 mmol/l; 6.5 mmol/l) it means that on the significance level of 5 % all hypotheses affirming that the average is either smaller that 5.1 mmol/l or greater than 6.5 mmol/l are denied. Our discoveries disproving the myth of accuracy of p values are the next argument for the use of confidence intervals.

For the sake of completeness we refer to the fact that when evaluating the diagnostic tests sophisticated methods are beginning to be used in greatest scale (Meloun et al, 2001; Gaško et al, 2001). After a century of "significance" it is expected that the essential statistical challenge is to develop new decision-making methods.

Exact p values in the same statistical tests calculated by diverse programmes can be different. In special cases the values can oscillate about 0.05.

Tab. 4. Greatest differences between p-values.

	t-test	Mann-Whitney test	F-test
Statistix vs Analyse-it	0	0.0181	0.0001
Statistix vs MedCalc	0	-	0.1834
Analyse-it vs MedCalc	0	-	0.1835

The acceptance of the zero hypothesis based on the value of p=0.049 or p=0.051 is clear nonsense, however, this approach dominates in the biomedical thinking (Henderson, 1993) wrongly based on the strictness of the set significance level.

In the metodological part of treatises not only the used statistical test but also the used statistical programme should be described as precisely as possible.

In the result part of treatises the exact p value should be stated without arbitrary decision on the resulting significance.

Confidence intervals provide more information than statistical significance tests. The best solution is to state the confidence interval as well as the exact p value in the treatises, however, if one of these two is to be omitted, it should be the p value. Confidence intervals are to be preferred to significance tests whenever they can be calculated*.

*Sources of support: The author was enabled to carry out some of his calculations at the Clinic of Neurology of the Teaching Hospital with Policlinic and the Faculty of Medicine, Safarikiensis University, Kosice.

References

Bland JM, Altman DG. Multiple significance tests: the Bonferroni method. Brit Med J 1995; 310: 170—172.

Chia KS. „Significant-itis" — an obsession with the P-value. Scand J Work Environment Health 1997; 23: 152—154.

Cleophas TJ, Zwinderman AH. Limitations of randomized clinical trials. Proposed alternative designs. Clin Chem Lab Med 2000; 37 (12): 1217—1223.

Evans SJW, Mills P, Dawson J. The end of the p value? Brit Heart J 1988; 60: 177—180.

Feinstein AR. P-values and confidence intervals: two sides of the same unsatisfactory coin. J Clin Epidemiol 1998; 51 (4): 355—360.

Gardner MJ, Altman DG. Confidence intervals rater than P values: estimation rather than hypothesis testing. Brit Med J 1986; 292: 746—750.

Gaško R, Klímová E, Balla B. Clinical Utilization of Multidimensional Statistical Methods in Slovak Medical Literature. Transactions of the Universities of Košice Biomedical subedition Folia Medica Cassoviensia 2001; 2: 61—66.

Goodman SN. Toward evidence-based medical statistics. I. The P value fallacy. Ann Intern Med 1999; 130: 995—1004.

Greenhalgh T. How to read a paper: Statistics for the non-statistician. I: Different types of data need different statistical tests. Brit Med J 1997; 315: 364—366.

Grimes DA, Schulz KF. An overview of clinical research: the lay of the land. Lancet 2002; 359: 57—61.

Haas T. Statistics. In: Kasal P, Svačina Š (Eds). Medical informatics (in Czech). Praha; Karolinum, 1998: 150—168.

Henderson R. Chemistry with confidence: Should Clinical Chemistry require confidence intervals for analytical and other data? Clin Chem 1993; 39 (6): 929—935.

Klímová E, Barlová E, Elbertová A, Čider V, Szilásiová J. Intracellular cytokines interferon gama and interleukin-4 produced with T lymphocyts of peripheral blood in patients with multiple sclerosis. Čes Slov Neurol Neurochir 2003; accepted to press.

Klímová E. Multiple sclerosis and interferons-beta. Rožňava: Roven Press, 2002: 4—64.

Lazurova I, Wagnerova H, Ladanyiova H et al. The impact of passive leg rising on sodium excretion in patients with liver cirrhosis. Bratisl Lek Listy 2001; 102 (4): 196—199.

Meloun M, Hill M, Cibula D. Exploratory biochemical data analysis: a comparison of two sample means and diagnostic displays. Clin Chem Lab Med 2001; 39 (3): 244—255.

Mikulecký M. How to calculate data, parameters and values. In: Hulín I, Mráz P (Eds). Introduction to scientific work in medicine. Bratislava: Univerzita Komenského, 1993: 114—150.

Perneger TV. What's wrong with Bonferroni adjustments. Brit Med J 1998; 316: 1236—1238.

Sterne JAC, Smith GD. Sifting the evidence — what's wrong with significance tests? Brit Med J 2001; 322: 226—231.

Received May 15, 2002.
Accepted January 27, 2003.