From the descriptive towards inferential statistics: Hundred years since conception of the Student's t-distribution

Peter G. Fedor-Freybergh [1] & Miroslav Mikulecký [2,3]
[1] Editor-in-Chief, Neuroendocrinology Letters
[2] 1st Internal Clinic, Medial Faculty, Comenius University, Bratislava, Slovak Republic
[3] Head, Dept. on Biometrics and Statistics, Neuroendocrinology Letters

Neuroendocrinol Lett 2005; **26**(3):167–172 **PMID:** 15990717 NEL260305E01

Introduction

It is crucially important to differentiate between the descriptive statistics on one side and inferential, inductive, mathematical statistics on the other. Only the latter is useful for obtaining the optimal information and for making effective decisions under conditions of uncertainty, because it defines probabilities of conclusions. That, of course, gains the real importance only in situations where the practical action is expected to be derived directly from statistical analysis, as in modern economy. In medicine, however, there are sometimes studies performed, where the only practical outcome is the obtaining of a scientific or paedagogic degree – so called "Mickey Mouse" studies [1]. Thus, we will be speaking about the fundamental orientation of scientific work in medicine.

Example for a mere descriptive statistics is the arithmetical mean with its standard deviation or standard error. That alone does not define any probability. In the contrast, the mean with the interval of 95% confidence yields a probable prediction for the underlying population on the basis of the sample. In other words, the inferential statistics attempts to address the relation between the sample of measurements and their corresponding (usually fictive) population which has been the sample drawn from, while the descriptive one does not. It ignores the difference between the sample characteristics, as for example the arithmetical mean $\boldsymbol{m}$ and standard deviation $\boldsymbol{s}$, and the population parameters, for example mean $\boldsymbol{\mu}$ and standard deviation $\boldsymbol{\sigma}$.

The importance of the difference between both approaches depends upon the sample size. The population parameters are – if the way of sampling was appropriate – sufficiently approximated with a sample size over 200. Theoretically,they are exactly valid for the infinite sample size. In this case, the Gaussian normal distribution is appropriate. The interval of 95% confidence will then be $\boldsymbol{m \pm 1.96.s}$. Nevertheless, under 200, and particularly under 30 measurements the difference between sample and population becomes more and more critical. Consequently, the use of mean and standard deviation, i. e. the Gaussian normal distribution for a few measurements is nonsense.

But exactly that was the statistical practice in XIX century! The normal distribution was glorified: it was found in remote heavenly objects the same way as in Quételet's measurements of the chest circumference of 5738 Scottish soldiers [2]. Ronald Aylmers Fisher [3] commented this abuse of the normal distribution by the words "...the traditional machinery of statistical processes." "Not only does ... take a cannon to shoot a sparrow 'but it misses the sparrow!".

With the industrial revolution, the attempts emerged to improve and organize production of goods on the basis of science. That means, to measure and evaluate variables important for the quality of the final product. Interestingly, that came the first time to the realization in production of quality beer [4]. In the Guinness brewery, for example, there has been found that the quantity of soft resin in the hops paralleled the qualitative assessment of hop "condition". Similarly, the malting quality of barley was shown to depend on nitrogen content. The brewers began routine measurements and accumulated data. Nevertheless, they ran into difficulties because their measurements varied. "They had no way of judging whether the differences they found were effects of treatment or accident. Two difficulties were confounded: the variation was high and the observations were few."(4). The brewers took these problems to their colleague – William Sealey Gosset (1876–1937). He had, for instance, to evaluate the barley experiments from four farms. He realized that the estimate ***m*** is influenced in an unknown way by the error in the estimate ***s***. Gosset posed therefore the following main question: how much wider should the error limits be to make allowance for the error introduced by using the estimates ***m*** and ***s*** instead of the parameters μ and σ? This was the ingenious principal question. But the leading person in biometrics of these times – Karl Pearson, one of the fathers of the Helmert-Pearson χ^2-distribution, did not know the answer!

Nevertheless, the answer, elaborated soon thereafter by the non-mathematician Gosset, became to the corner stone of the new, *inductive* uses of mathematics. New era in statistical evaluation of observations and measurements started just to be born.

The story was as follows. In July 1905, exactly hundred years ago, Gosset cycled during his summer holidays 20 miles to visit Pearson in his summer home in England. Gosset recalled later that "Pearson was able in about half an hour to put me in the way of learning the practice of nearly all the methods then in use".

During 1906–1907 spent Gosset a year at Pearson's laboratory in London and worked out the exact answer to his question. He was now able to calculate the values of the probable error of the mean using his seemingly simple criterion ***z*** (today Student's random variable ***t*** of the Student distribution)

$$(z =)\ t = (m-\mu) / s$$

(formula 1).

He tabulated the probability values of his criterion at first for samples of N = 2, 3, .., 10, i.e. – for univariate samples – for 1 to 9 degrees of freedom. And he used this theory immediately for solving a practical task: to identify the best barley on the basis of the Guinness barley experiments. Gosset concluded that Archer barley was the best one for Ireland. And Guinness were delighted. The *decision* followed, to buy all pure-line Danish Archer seed available to buy and to grow it all over the Irish island. The final result was the excellent quality of Guinness living beer, able to be transported without conservation means and safely, without any damage to its quality, to remote parts of the world. The name of Guinness became world famous. Guinness brewery was some time the biggest one in the world.

Gosset published his paper [5] on the probable error of the mean the next year (1908). Guinness did permit the publication but Gosset's name had not to appear. He could choose between "Pupil" or "Student". Gosset took the pseudonym Student.

Let us show in a very simplified manner the mathematical quintessence of Student's discovery in some details.

Let us start with the normal distribution. The random variable of the standardized normal distribution [6] is

$$c = (m-\mu) / \sigma$$

(formula 2).

In other words, the random deviation of the sample mean ***m*** from the population mean ***μ*** is expressed by its standard deviations ***σ***, ***μ*** and ***σ*** alone being given in the units of original measurements, e.g. meters. It follows that the mean of the standardized normal distribution is 0 and its standard devitation is 1. The graphic display of the corresponding probability density function is the well known Gaussian curve. Its integral, *i.e.* the area under the curve from $-\infty$ to $+\infty$ is equal to the probability 1. The values of ***c***, expressed with the aid of ***σ*** as the unit, from −1 to +1 demarcate the probability of 0.68 while the ***c*** values from −1.96 to +1.96 demarcate the probability of 0.95. This is exactly valid for infinite number of measurements. Analogical values for example for 4 measurements, as in the Gosset's problem, will be for the ***t*** (instead of ***c***) values from −1 to +1 the probability 0.609 and for the probability of 0.95 the ***t*** values from −3.1824 to +3.1824. While the ***c*** values are expressed with the aid of the population standard deviation ***σ*** as the unit (formula 2), the ***t*** values are expressed with the aid of the sample standard deviation ***s*** (formula 1)**.** Practically, in the case of 4 measurements, their sample mean can randomly fluctuate with 95% probability by 3.1824.***s*** up and down, not by 1.96.***s*** what is the case for at least 200 measurements, exactly for their infinite number when ***s*** becomes to ***σ*** .

It remains to explain how the random variable χ^2 , describing the random fluctuation of the sample variance, was involved in the Gosset's July bike expedition. The random variable $\boldsymbol{\chi^2}$ is derived from the same standardized Gaussian curve as the variable ***c***, describing the random fluctuation of mean, just the squares $\boldsymbol{c^2}$ are taken and summed up:

$$\chi^2 = c_1^2 + \ldots + c_{n-1}^2 = (x_1-\mu)^2/\sigma^2 + \ldots + (x_{n-1}-\mu)^2/\sigma^2 = (n-1).s^2/\sigma^2$$

(formula 3).

From that it follows

$$\chi^2 / (n-1) = s^2 / \sigma^2$$

(formula 4).

In other words, this random variable describes the random fluctuation of the sample variance around the population variance as standardized by the degrees of freedom n−1.

Going back to the Student's random variable ***t***, we get by combining formula 1 and 4

$$t = c. \quad (n-1) / \quad \chi^2$$

(formula 5).

It is evident that the Student random variable ***t*** is a function of the random variable of the standardized normal distribution ***c***, of the random variable $\boldsymbol{\chi^2}$ and of the degrees of freedom N−1, i. e. the number of measurements diminished by 1.

It has to be stressed that, in fact, the calculations of the critical values of these distributions are not so simple as it would seem from the simple formulae. The mutually stochastically *independent* realizations of the separate random variables, e.g. of ***c*** and $\boldsymbol{\chi^2}$, have to be taken into account. Only in formula 3 where the χ^2 variable is the function of the sole variable ***c***, is the critical value of $\boldsymbol{\chi^2}$ for the probability 1−0.95=0.05 equal to the squared critical value ***c*** for the same probability of standardized normal distribution: $3.841 = 1.96^2$. Both these values are abundantly used in many statistical calculations, for contingency tables with qualitative variables as $\boldsymbol{\chi^2}$ and for continual variables as ***c***. It would be quite incorrect to 'calculate" the critical value of ***t*** at 2P = 0.05 for 2 measurements (degrees of freedom equal 1) from the critical values of ***c*** and $\boldsymbol{\chi^2}$ at 2P = 0.05 and one degree of freedom using the formula 5: the "resulting value" of ***t*** would then be 1.96 . (2−1) / 3.841 = 1 while the correct table value is ***t*** = 12.7062.

It is important, from the paedagogical point of view, to explain the random variables in the same sequence as they logically and historically originated: ***c*** **(=u),** $\boldsymbol{\chi^2}$**,** ***t*** [7, 8, 9], not in the sequence ***c*** **(=Z),** ***t*** **,** $\boldsymbol{\chi^2}$ [10, 11, 12, 13] which, otherwise, would be reasonable on the basis of the close rela-

tion between normal and ***t***-distribution: the normal distribution, in fact, after the Student's discovery "degenerated" to a special case of the ***t***-distribution for infinite number of degrees of freedom. Besides, the evolution of the normal distribution from the binomial one has to be explained and understood, too.

The Student ***t***-distribution and the Student ***t***-test are nowadays one of the basic tools of data processing in science as well as in practice. In some sense, Student belongs to the mostly cited authors ever. Thus, our OLDMEDLINE and MEDLINE SEARCH has shown that since the year 1950 up to April 4, 2005, there are 19 313 citations of "Student's distribution", "Student's test", "***t***-distribution" or "***t***-test". For comparison, R.A Fisher, one of the most respected statisticians ever, has altogether 1398 items in the same time span.

Despite of the high citation rate of the Student's method, the spirit of his legacy remains widely misunderstood. The random variable ***t*** is used predominantly for performing the Student *test*. It has, however, more appropriate application for calculating the interval *estimates*. Thus, the confidence intervals are given simply as ***t***-multiple of standard error. Already R.A. Fisher preferred estimates before hypothesis testing [14].

This policy is recently widely supported by the British Medical Journal Group [15,16]. The target of research is often to test the presence of an *effect*. The P value, understood appropriately, gives only a dichotomic, black or white answer – yes or not – while confidence interval for an effect predicts the probable range of an effect size in corresponding population as expressed in the original metrics, i.e. for instance as blood pressure values.

It is hardly understandable why so many authors and Journals perseverate in ignoring the confidence estimates in favour of standard deviations or errors and of tests. For example,The Lancet published an article in favour of confidence intervals, too [17]. Fourteen years later, however, a large international study of drug effect [18] presented means, standard errors (for sufficiently large sample sizes) and P values, calculated obviously with the aid of Student' s ***t***-test. In one case, however, the claimed P value (0.02), meaning statistically significant effect, should be corrected to 0.07, i.e. nonsignificance.

One especially very often used and in XXI century hardly acceptable "routine" way of presenting a research paper is to give the sample mean and standard deviation or standard error (or even "+ –" without specification) before some maneuvre and the same after it, and to express the difference by P value. Morever, this happens sometimes for low or even very low sample sizes! That is exactly the way of overusing the Gaussian distribution, known from the XIX century.

This practice ignores the one century old Gosset's guess that normal distribution is not appropriate for low sample sizes. Their effect on the probability of error is demonstrated by the following table. The calculations were performed with the aid of Texas Instruments Programmable 58 calculator.

Degrees of freedom n–1	Probability within the interval of ***t*** <–1;+1 >	Interval of ***t*** including the probability of 0.95
1	0.500	<–12.7062;+12.7062 >
2	0.577	<–4.3027;+4.3027 >
3*	0.609	<–3.1824;+3.1824>
5	0.637	<– 2.5706;+2.5706>
10	0.659	<– 2.2281;+2.2281>
20	0.671	<–2.0860;+2.0860
100	0.680	<–1.9840;+1.9840
∞	0.683	<–1.9600;+1.9600>

*Gosset's original case

The second column of the table demonstrates that one standard deviation for different degrees of freedom includes different probabilities. Consequently, it is a flawed statistical characteristics. The third column shows that the error range for lower sample sizes is wider.

If the ***t*** values are multiplicated by the standard error expressed in the original metrics (e.g. kilograms), the confidence interval predicting the random fluctuation of mean, to estimate the location of the population mean μ , is obtained. If instead of the standard error the standard deviation is used (more exactly, the standard error multiplicated by n+1, the tolerance interval for one individual will be obtained [19]. The latter is usually much wider and therefore hardly acceptable for those who need to present clearly "significant" effects . Tolerance, however, is – versus confidence – more appropriate for clinical practice where one patient should stand in the limelight of the attention. Insufficient, too rare use of tolerance limits is criticized [20]: they have "enormous importance in medicine and biology for setting normal ranges. So far, they were only rarely exactly calculated according to rules for tolerance limits." That happens despite of relative simplicity of these calculations.

Nowadays, estimates are calculated on the basis of estimation theory for much more complicated situations than univariate sample of data. Corresponding mathematical research was performed also in Slovak Republic and presented internationally [21, 22].

What consequences could be drawn for our Journal?

At first, any author should check his paper, already published, from the point of view of principles outlined in this Editorial. It would be interesting to overcalculate some results on the basis of estimates if there were not given.

As an example, let us look in details on the last issue of the year 2004. The majority of papers gives, indeed, often standard deviation for low to very low sample sizes, as 4, 5, 7, 8, 9. In the formulation of conclusions, the P values dominate. Mathematically, the calculations are, of course, correct. Nevertheless, the outcome is usually not optimal from the point of clear information and decision making in practice.

At second, for the future, the present Editorial intends to change the statistical style of presentation of results in the manner, compatible with the progress of statistics in the XX century. The statistical methods, recommended in the present instructions for authors, are chosen appropriately, in agreement with the principles proclamated in this Editorial. They, however, should be extended in more details.

We conclude with a challenge. Let us leave the XIX century also in the statistical processing of our data in research papers! At least, for the very beginning, let us abandon using the standard deviation or error for small samples! Let us understand and follow the legacy of the ingenious brewer from Dublin created one century ago.

Acknowledgements

We are indebted to Dipl. Ing. Irina Fialková, Slovak Medical Library, Bratislava, for the corresponding citation analysis.

REFERENCES

1 Altman DG, Bland JM. Improving doctors´ understanding of statistics. J R Statist Soc A 1991; **154**/2: 223–267.
2 Swoboda H. Knaurs Buch der modernen Statistik. Muenchen-Zuerich: Droemer Knaur Verlag. 1971. Czech translation: Normální rozdělení. Dějiny normální křivky. S.73–80 in: Moderní statistika. Prag: Verlag Svoboda. 1977. 352 S.
3 Fisher R.A. Statistical methods for research workers.12th ed. Edinburgh 1954.
4 Fisher Box J. Guinness, Gosset, Fisher, and small samples. Stat Sci 1987; **2**:45–52.
5 Student. The probable error of a mean. Biometrika 1908; **6**:1–25.
6 Diem K,Seldrup J. The standardized normal distribution. P.198 in: C. Lentner (ed), Geigy scientific tables. Vol.2. Introduction to statistics. Statistical tables. Mathematical formulae. 8th ed. CIBA-GEIGY, Basle 1982.

7 Mikulecký M. The fundamentals of biometry for experimental and clinical medicine. (In Slovak.) Comenius University Bratislava. 1985.208 pp. Reviewed by F Halberg, Chronobiologia.
8 Weber E. Die c2 Verteilung von FR Helmert und Karl Pearson. S. 152–162 in: Weber E. Grundriss der biologischen Statistik. 6. Aufl. Jena: VEB Gustav Fischer Verlag. 1967. 674 S.
9 Weber E. Die t-Verteilung von STUDENT. S.162–164 in: Weber E. Grundriss der biologischen Statistik.6. Aufl. Jena: VEB Gustav Fischer Verlag. 1967. 674 S.
10 Sachs L. Die Student-Verteilung. S.109–110 in: Sachs L. Angewandte Statistik. Anwendung statistischer Methoden. 6.Aufl. Berlin, Heidelberg, New York, Tokyo: Springer-Verlag. 1984. 552 S.
11 Sachs L. Die χ2 – Verteilung. S.110–115 in: Sachs L. Angewandte Statistik. Anwendung statistischer Methoden. 6.Aufl. Berlin, Heidelberg, New York, Tokyo: Springer-Verlag. 1984. 552 S.
12 Sokal RR, Rohlf FJ. Student's t-distribution. P.145–147 in: Sokal RR, Rohlf FJ. Biometry. The principles and practice of statistics in biological research. 2nd ed. San Francisco: W.H. Freeman Co. 1981, 859 pp.
13 Sokal RR, Rohlf FJ. The chi-square distribution. P.152–155 in: Sokal RR, Rohlf FJ. Biometry. The principles and practice of statistics in biological research. 2nd ed. San Francisco: W.H.Freeman Co. 1981, 859 pp.
14 Brown GW. Errors, Type I and II. Am J Dis Child 1983; **137**:586–591.
15 Gardner MJ, Altman DG. Confidence intervals rather than P values; estimates rather than hypothesis testing. Br Med J 1986; **282**:746–750.
16 Gardner MJ, Altman DG. Confidence intervals rather than P values; estimates rather than hypothesis testing. P. 6–19 in: Gardner MJ, Altman DG. Statistics with confidence. Confidence intervals and statistical guidelines. London: Br Med J Publishing House. 1990. 140 pp.
17 Bulpitt JC. Confidence intervals. The Lancet 1987; February 28, i: 494–497.
18 Investigators of the Diabetes Atherosclerosis Intervention Study. Effect of fenofibrate on progression of coronary artery disease in type 2 diabetes. The Lancet 2001; **357**:905–910.
19 Mikulecký M. Confidence and tolerance intervals – a tool for biomedical data analysis aimed at clear evidence. Kardiológia / Cardiology 2004; **13**:211–215.
20 Anonym. Toleranzgrenzen stetiger Verteilungen. S.192 in: Wissenschaftliche Tabellen Geigy. Teilband Statistik. Logarithmen. Zahlentafeln zur Statistik. Mathematische Symbole, Definitionen und Formeln. Einfuehrung in die Statistik. 8. Auflage. CIBA-GEIGY Ltd. Basle 1980.
21 Kubáček L. Foundations of estimation theory. Amsterdam, Oxford, New York,Tokyo: Elsevier. 1988.
22 Kubáčková L. Foundation of experimental data analysis. Boca Raton, Ann Arbor, London, Tokyo: CRC Press. 1992.