Available online at www.sciencedirect.com

ScienceDirect

Journal of Chromatography B, 847 (2007) 305–308

JOURNAL OF CHROMATOGRAPHY B

www.elsevier.com/locate/chromb

Discussion

Statistical power and analytical quantification

Pedro Araujo*, Livar Frøyland

National Institute of Nutrition and Seafood Research (NIFES), P.O. Box 2029 Nordnes, N-5817 Bergen, Norway

Received 26 April 2006; accepted 7 October 2006
Available online 27 October 2006

Abstract

It is suggested that power analysis should be formally incorporated into quantification experiment reports in order to substantiate the conclusions derived from experimental data more effectively. The article addressed the issues of power analysis calculation, sample size estimation and appropriate data reporting in quantitative analytical comparisons. Illustrative examples from the literature are used to show how the described power analysis theory could be applied in practice.

Keywords: Statistical power; Analytical comparisons; Quantification; Validation; Chromatographic methods

1. Introduction

Analytical quantification comprises a wide variety of experimental procedures and well-established instrumental techniques which can be used in conjunction to tackle a particular problem. The multiplicity of available procedures and instrumental techniques also provides an opportunity for comparison and learning about the use, merits and limits of individual instruments. Although, comparison studies are always useful to propose new and more efficient methodologies in order to enhance the accuracy and quality of the results, save time, efforts and resources, an appropriate statistical analysis must always support the results of the comparisons. Failure to comply with this premise can have serious implications in basic and applied sciences.

Quantification experiments depend on statistical inference and should be structured around a hypothesis which suggests, for instance, that there are not real differences between the concentration levels of a particular contaminant in drinking water measured by an approved and an alternative instrumental technique. The idea of cancelling out the difference between the two instrumental techniques by assuming no statistical significance is called the null hypothesis (H_0). There is no method to determine whether the status of H_0 is, in fact, true or false. Any decision about rejecting or accepting H_0 on the basis of a statistical test is always accompanied by some uncertainty due to the inherent random error of the experimental data [1,2]. Fig. 1 illustrates that the probabilities to make a right decision after performing an appropriate statistical analysis are $1-\alpha$ and $1-\beta$ (clockwise direction) also known as confidence level and statistical power, respectively. Conversely, the chances to make a wrong judgment are α and β (anticlockwise direction) also known as Type I and Type II errors, respectively. The probability that H_0 is rejected even though it is true (Type I error) is computed by quoting the significant level α of the test ($\alpha = 5\%$ is generally reported in analytical comparisons), while the probability that H_0 is accepted even though it is false (Type II error) is often not even considered. The reason for this omission could lay behind the fact that most statistical books provide a cursory treatment of the subject. Besides, a Type II error is more difficult to quantify since it requires an investigation of the power of the test. This paper aims to promote understanding of statistical power analysis by outlining its conceptual basis in the context of analytical chemistry quantification.

2. Statistical power

Statistical power measures the confidence with which it is possible to detect a particular difference or effect if one exists and it is generally defined as the probability of not committing a Type II error. If power is not high enough, in quantification experiments aiming at comparing various analytical methodologies, it is possible then to conclude wrongly that the compared methods yield the same results which in turn can have serious

* Corresponding author. Tel.: +47 55905115; fax: +47 55905299.
E-mail address: pedro.araujo@nifes.no (P. Araujo).

1570-0232/$ – see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.jchromb.2006.10.002

implications, for instance in the analysis of toxic compounds in food products for human consumption.

The statistical power of a comparative quantification study is defined by the number of replicates analyzed, the level of significance α, the overall variance s_o^2 and the non-central parameter λ which measures departures from the null hypothesis H_0 [2,3].

2.1. Power calculation in quantitative analytical comparisons

Consider the determination of a compound x in a determined matrix by using I different instrumental techniques and involving J_i sample replicates per instrumental technique at a 5% significant level and on the assumptions that data normality, homoscedasticity and independency of residuals are met. The non-central parameter λ is calculated by estimating the variances within (s_o^2) and between (s_b^2) the different instrumental techniques as follows:

$$s_o^2 = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J_i}\left(x_{ij}-\bar{\bar{x}}\right)^2}{\sum_{i=1}^{I}(J_i-1)} \tag{1}$$

$$s_b^2 = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J_i}J_i\left(\bar{x}_i-\bar{\bar{x}}\right)^2}{I-1} \tag{2}$$

$$F_{\text{exp}} = \frac{s_b^2}{s_o^2} \tag{3}$$

$$\lambda = (I-1)\times F_{\text{exp}} \tag{4}$$

The term x_{ij} in Eq. (1) represents the concentration of replicate j determined on the instrument i, the terms $\bar{\bar{x}}$ and $\bar{x}_i$ in Eqs. (1) and (2) represent the overall average concentration and the average concentration at each instrumental technique, respectively, and the term F_{exp} in Eqs. (3) and (4) is the experimental Fisher ratio.

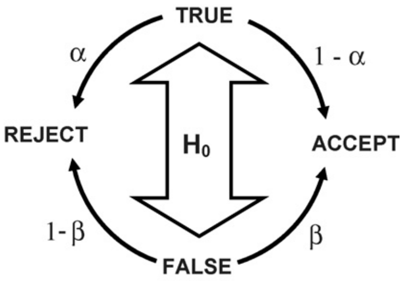

Fig. 1. Representation of the main concepts discussed in this article. Clockwise direction: confidence level $(1-\alpha)$ and statistical power $(1-\beta)$. Anticlockwise direction: Type I error (α) and Type II error (β).

The average concentrations $\bar{\bar{x}}$ and $\bar{x}_i$ in Eqs. (1) and (2) are defined by the expressions:

$$\bar{x}_i = \frac{\sum_{j=1}^{J}x_{ij}}{J_i} \tag{5}$$

$$\bar{\bar{x}} = \frac{\sum_{i=1}^{I}\bar{x}_i}{I} \tag{6}$$

Eq. (1) requires the availability of the whole experimental data (every x_{ij} measured), which is not always possible, especially if a prospective or a retrospective analysis is based on information gathered from published works where in general the information provided is limited to the number of replicates, mean values and their standard deviations. In such cases s_o^2 is calculated by the expression:

$$s_o^2 = \frac{\sum_{i=1}^{I}\left[(J_i-1)\times\sigma_i^2\right]}{\sum_{j=1}^{J_i}(J_i-1)} \tag{7}$$

It is possible to determine the power of a particular comparative quantification experiments by substituting the values of I, J_i and λ described above in the Laubscher's square root normal approximation of non-central F-distribution [4] which in the context of this article becomes:

$$z_{1-\beta} = \frac{\left[2\sum_{i=1}^{I}(J_i-1)-1\right]^{1/2}\left(\left((I-1)/\sum_{i=1}^{I}(J_i-1)\right)F\right)^{1/2}-\left[2(I-1+\lambda)-\frac{I-1+2\lambda}{I-1+\lambda}\right]^{1/2}}{\left[\left((I-1)/\sum_{i=1}^{I}(J_i-1)\right)F+(I-1+2\lambda)/(I-1+\lambda)\right]^{1/2}} \tag{8}$$

The unit normal percentile value for power $z_{1-\beta}$ and the tabulated Fisher ratio F with $I-1$ and $\sum_{i=1}^{I}(J_i-1)$ degrees of freedom in the numerator and denominator, respectively, are computed from reported statistical tables [5]. Alternatively, the reader can determine the power by consulting any available online-based Laubscher's non-central F-distribution calculator [6].

The relationship between Eqs. (4) and (8) is one of the most important conceptual issues in power analysis. It implies that null hypothesis always means that the λ is zero [2].

Although the authors of the present article have personally used or reviewed some of the commercial software packages it is not their intention to present a comprehensive list of packages but instead to highlight the underlying principles involved in power analysis and method comparison. The interested reader is referred to the exhaustive review of Thomas and Krebs [7] who compared 29 popular commercial software in terms of cost, operating systems, easy to use, easy to learn, calculation methods, power and sample size capabilities, z-test, t-test, fixed and random effects ANOVA, repeated measurements, regression, correlation, non-parametric test, probability calculator, etc.

2.2. Replication and data reporting

One important aspect of power analysis at the design stage of a study is the selection of the number of replicates. It is generally accepted that an increase in the number of replicates provides a higher return in terms of power. However, the relationship between sample size and power is not linear and consequently at some specific α levels, a huge increase of the replicates brings about only a modest increment in power. Various approaches have been reported to determine an appropriate number of replicates [8,9]. Some of these approaches share common features that have been used to generate rules of thumb which are useful in determining the number of replicates necessary to give a high power. For calculating the dependence of a required number of replicates on the statistical parameters, the following expression can be used:

$$J_i \approx \left(z_{\alpha/2} + z_\beta\right)^2 \tag{9}$$

It is advisable to build a power table in cases where the number of instrumental techniques is known in advance by substituting in Eq. (8) different number of replicates (J_i) and several representative non-central parameter λ estimated from pilot studies. This, then, enables the analyst to select a sensible number of replicates and a suitable statistical power.

It is important to mention that there is not a conventional criterion to determine what is a suitable statistical power, however a value of 80% is generally considered the minimum desirable.

Data reporting is a critical aspect of power analysis and sometimes is perceived as less important than the experimental conditions and data collection description. Although there are different ways to report data from comparative quantification studies, a standard report should always contain information on three parameters, namely, number of replicates, averages and standard deviations. By reporting these parameters the interested readers can perform a retrospective power analysis in order to assess if the number of replicas used was adequate and the power of the analysis sufficient to reach the statistical conclusions derived from a particular study. In addition, such parameters enable conducting a prospective power analysis in a design phase of an intended comparative study and consequently determining a rational number of experimental replicates per instrument, necessary to empower a future study. It is important to highlight that a practical guide for analytical method validation with a description of a set of minimum requirements for a method [10], based on the United States Pharmacopeia, the International Conference on Harmonisation and the Food and Drug Administration, has established that any data report should use at least three replicates to judge statistically the acceptability of an analytical method.

Accepting the validity of reported analytical comparisons without the above mentioned parameters and considerations could pose a serious threat to public health especially in studies aiming at comparing a certified method against a new one used, for instance in clinical, food, water or beverage analysis.

3. Illustrative examples of power analysis in reported quantitative analytical comparisons

The application of the power analysis theory described above is demonstrated in the analysis of two published studies.

3.1. Example 1

Prospective and retrospective power analysis is of leading importance to substantiate the conclusions derived from quantification experiments more effectively. An inspection of the literature revealed that in general, the emphasis of statistics in various quantification analysis has been on evaluating the probability that the null hypothesis will be rejected when it is true (α error or Type I error). For instance, in a study aimed at determining cholesterol in milk fat by using the internal standard technique [11], two instrumental methods, supercritical fluid chromatography and gas chromatography, were compared and the following concentrations of cholesterol 3.08 ± 0.0089 and 3.13 ± 0.0222 in milligrams per gram were reported for triplicate samples, respectively (the coefficient of variations were the statistical values reported by the authors which have been transformed into standard deviations for the explanation of the present example). The authors concluded that both techniques were suitable for the intended analysis at a significant α level of 5%. By using the reported data and Eqs. (7), (2) and (4) described above the values 4.29×10^{-4}, 3.75×10^{-3} and 8.74 for s_o^2, s_b^2 and λ were calculated, respectively. By substituting an F-tabulated value of 7.709 (1 and 4 degrees of freedom in the numerator and the denominator, respectively), in Eq. (8) a power of 60% is estimated. This statistical result implies that the assertion of no difference between both techniques would have a 40% chance of being wrong ($\beta = 0.40$). We incline towards the 40% chance of being wrong after averaging the standard deviations reported by these authors (Tables 2 and 3 of the published article) and obtaining coefficient of variations for the standard deviations of approximately 140% in both techniques. It can be demonstrated that the authors of this reported work should have increased the number of replicates from three to eight in order to reach a power of 80% by substituting $z_{0.05/2} = 1.96$ (95% confidence level) and $z_{0.20} = 0.84$ (80% statistical power) in Eq. (9) as follows:

$$J_i \approx (1.96 + 0.84)^2 \approx 7.84 \approx 8$$

By increasing the power from 80 to 95% an increase of 65% in the number of replicates is estimated (12.96 replicates estimated by using the values $z_{0.05/2} = 1.96$ and $z_{0.05} = 1.64$). Similarly, by using the previous estimated λ (8.74), substituting different replicate (J_i) and F-tabulated values in Eq. (8) it is possible to construct a power table that may help in the selection of a sensible number of replicates and an appropriate statistical power. Table 1 shows that by changing the values of J_i from 3 to 10 in Eq. (8) a total number of nine replicates seems advisable to reach a power of 80%. We are not going to speculate on the implications of this article; however, we must remember that without sufficient statistical power, data-based conclusions may be useless and sometimes the consequences of such conclusions could result in the implementation of inappropriate actions.

3.2. Example 2

To illustrate the importance of reporting, a study aimed at comparing different methods for the determination of

Table 1
Statistical power table constructed by increasing the number of replicates in Eq. (8)

Instrumental techniques	I	2							
Sample replicates	J_i	3	4	5	6	7	8	9	10
	$I-1$	1							
Degrees of freedom	$\sum_{i=1}^{2}(J_i-1)$	4	6	8	10	12	14	16	18
Non-centrality parameter[a]	λ	8.740							
Fisher tabulated ratio	F	7.709	5.987	5.318	4.965	4.747	4.600	4.494	4.414
Statistical power (%)	$1-\beta$	60	70	74	77	78	79	80	80

[a] Estimated from reference [11].

ochratoxin-A (a mycotoxin involved in kidney damage and potentially carcinogenic for humans) in wine is discussed [12]. The article gives a careful account of the accepted immunoaffinity liquid chromatography with fluorescence detection (IA–LC–FL) method and the alternative reverse phase octadecylsilica solid phase extraction liquid chromatography tandem mass spectrometry (RP18-SPE–LCMS/MS) method used for the intended purpose along with a detailed description of the data collection. The data report contains only several ochratoxin-A averages estimated by using duplicate wine samples from different European regions. The conclusion derived from the data report was that the RP18-SPE–LCMS/MS method represents a genuine alternative to the already established, effective and accepted IA–LC–FL. Unfortunately the authors did not consider the acceptable minimum of replicates in their report to judge the validity of RP18-SPE–LCMS/MS as a reliable alternative. In addition, by using the values 0.59 ± 0.07 and 0.53 ± 0.04 μg/ml of ochratoxin-A for the Austrian wine (the authors actually reported the standard deviation of the slopes) by SPE–LCMS/MS and IA–LC–FL, respectively, and following the same line of reasoning used in the previous example, we concluded that the allegation of no difference between RP18-SPE–LCMS/MS and IA–LC–FL would have a 89% chance of being wrong. It should be noted that we are not judging the veracity of the conclusions reached by these authors, our sole reason for discussing this reported study is to demonstrate that comparisons without adequate data reports or statistical analyses are futile and could in the worst case represent a serious health hazard. Researchers must be aware that whenever a comparison study is undertaken critical readers will be interested in testing their findings.

4. Conclusions

The benefits of power analysis, sample size estimation and appropriate data reporting in quantitative analytical comparisons have been demonstrated. Applied researchers should use power analysis where possible to derive more reliable conclusions from their findings and to enhance the quality of their research. Editors and reviewers of scientific journals could promote the application of power analysis tools by requiring power estimates from authors submitting articles, especially in cases where the implementation of a new methodology could have serious implications in human health.

References

[1] J.H. Zar, Biostatistical Analysis, Prentice Hall Inc., Englewood Cliff, NJ, 1984, p. 699.
[2] J. Cohen, Statistical Power for the Behavioral Sciences, Lawrence Erlbaum Associates Inc., Hillsdale, NJ, 1988, p. 469.
[3] http://www.for.gov.bc.ca/hre/forprod/pwrwksp.pdf.
[4] N.F. Laubscher, Ann. Math. Stat. 1 (1960) 1112.
[5] D.V. Lindley, W.F. Scott, New Cambridge Statistical Tables, Cambridge University Press, Cambridge, 1996.
[6] http://www.danielsoper.com/statcalc/calc06.aspx.
[7] L. Thomas, C.J. Krebs, Bull. Ecol. Soc. Am. 78 (1997) 126.
[8] P. Feigl, Biometrics 34 (1978) 111.
[9] W.G. Cochran, G.M. Cox, Experimental Designs, John Wiley & Sons, London, 1953, p. 23.
[10] J.M. Green, Anal. Chem. 68 (1996) 305A.
[11] W. Huber, A. Molero, C. Pereyra, E.M. de la Ossa, J. Chromatogr. A 715 (1995) 333.
[12] A. Leitner, P. Zöllner, A. Paolillo, J. Stroka, A. Papadopoulou-Bouraoui, S. Jaborek, E. Anklam, W. Lindner, Anal. Chim. Acta 453 (2002) 33.