A new approach to sample size calculations for the power of testing and estimating population means of gaussian distributed variables

Bruno Mario Cesana[1], Paolo Antonelli[2]

Section of Medical Statistics and Biometry,
Department of Biomedical Sciences and Biotechnologies,
Faculty of Medicine and Surgery, University of Brescia
[1]Associate Professor of Medical Statistics
[2] Contract Professor of Medical Statistics

Corresponding Author:
Cesana Bruno Mario
Sezione di Statistica Medica e Biometria, Dipartimento di Scienze Biomediche e Biotecnologie,
Facoltà di Medicina e Chirurgia,Università degli Studi di Brescia,
V.le Europa 11, 25123 Brescia, Italy
e-mail address: cesana@med.unibs.it
Phone: #39-030-3717530 - Fax: #39-030-3717488

Summary
Sample size calculations for demonstrating and estimating an effect are two distinct approaches that respectively concentrate on the power of the statistical test and on the precision of the confidence interval: however, they both have some unsatisfactory aspects.
This paper describes a sample size calculation method that, starting from the precision given by the power-based sample size, is based on the probability (conditional on the coverage) of obtaining both a statistically significant result (the power of the statistical test) and sample half-width confidence intervals lower than a considered precision threshold (power of the confidence interval). This procedure, therefore combines the two powers of the inferential statistical procedures, fulfils the regulatory requirements and the needs of a scientific research that are mainly focused on the probability of demonstrating a clinically relevant difference, and demands a generally acceptable increase of the power-based sample size.
The proposed method is pertinent to sample size calculation for controlled clinical trials planned to compare a new treatment against a reference one (or a placebo) on Gaussian distributed variables according to crossover or parallel group designs.

KEY WORDS: *confidence interval expected width, confidence interval power, controlled clinical trials, statistical test power.*

INTRODUCTION

A recently debated subject is whether sample sizes should to be calculated on the basis of the power for demonstrating a relevant effect such as a standardized difference (δ/σ) or effect size (1), or on the basis of the precision of the effect estimate such as a required width or half-width of the confidence interval (usually standardized) according to McHugh and Le (2), Greenland (3), Bristol (4), Goodman and Berlin (5), Borenstein (6), Nicewander and Price (7), Feinstein and Concato (8), and Hahn and Meeker (9).

The latter approach is strongly supported by the fact that, unlike the statistical significance of a test, a confidence interval (CI) precisely characterizes the magnitude of the effect. On the other hand, confidence interval-based sample sizes give a probability of obtaining the required precision of about 0.50, and, in addition, allow a power of about 0.50 in demonstrating an effect size (δ/σ) equated to the standardized CI half-width chosen as a precision threshold.
Kupper and Hafner (10), Grieve (11), and Beal (12) proposed to augment the sample sizes calculated on the basis of a CI required width (or half-width)

© CIC Edizioni Internazionali

for increasing the probability that the sample confidence interval widths (or half-widths) are less than the required precision threshold, expressed as a function of estimated standard deviation units; in addition, Beal (12) defined this probability as the CI power and considered the sample size calculation conditionally on the fact that the confidence interval includes the true value of the parameter of interest.

Grieve (13) gave an alternative calculation method without the numerical problems noted by Beal (12) and considered the uncertainty about the variance by incorporating it through a prior distribution. Finally, Wang and Kupper (14) extended this CI-based sample size calculation approach to the case of the heteroskedasticity of two normally distributed populations, unconditionally and conditionally on the coverage.

However, following the above suggestions, only the sample sizes for the highest values of the CI power, in the case of standardized CI half-widths > 1.20, give a satisfactory test power to demonstrate the same value of the effect size, as it can be immediately seen from the tables produced by Machin et al. (15); therefore, power values are unsatisfactory for demonstrating a small (0.2), medium (0.5) or even large effect size (0.8) according to Cohen (1).

Finally, Jiroutek et al. (16) have proposed a new method for calculating sample sizes for CI-based inferences that, starting from a required width, simultaneously considers width, validity (term used to describe whether or not "the parameter of interest is contained between the observed upper and lower confidence interval bounds") and rejection in an attempt to answer what they defined "the clinician's best question": "given validity, how many subjects are needed to have a high probability of producing a confidence interval that correctly does not contain the null value when the null hypothesis is false and has a width not greater than the required width?". However, a calculation approach focused on a high precision can give sample sizes much greater than the power-based ones (as the Table 2 on page 584 of the Jiroutek et al. (16) paper shows), thus leading to substantial doubts concerning the real feasibility of the study and, consequently, the practical usefulness of the Jiroutek et al. (16) method.

So, it seems that the sample size calculation method proposed by the statistical literature for considering together the precision of an effect estimate and for demonstrating an effect has to rely on an approach based on a CI required width; however, this approach is not consistent with the regulatory ICH E9 Guideline (17) in which it has been stated that "the usual method for determining the appropriate sample size has to be based on the statistical test for the primary analysis". Indeed, it has to be stressed that it is unethical to enroll in controlled clinical trials more patients than the necessary sample size for demonstrating a clinically relevant difference with a satisfactory statistical test power (0.80-0.90); accordingly, it is not appropriate to consider sample size calculation for a required width if the obtained sample sizes give unsatisfactory statistical power values, or if they are too much greater than the power-based values for demonstrating a clinically relevant alternative hypothesis.

In agreement with the above outlined ethical point which is also very important for the actual feasibility of a study, Cesana et al. (18) and Cesana (19) considered respectively the comparison of a proportion against a reference value and the comparison of two paired and unpaired proportions and devised a sample size calculation approach based on a "two-step" procedure. Particularly, starting from the expected width of the CI under H_A (EW), given the power-based calculated sample size, and the probability of obtaining confidence intervals whose width is less than the expected one (PW_{LE}), they proposed to iteratively increase the power-based sample size until PW_{LE} reaches a prefixed satisfactory value without conditioning on the coverage.

So, the increased sample size, obtained at the iterative second step of the "two-step" procedure, allowed obtaining the power of the statistical test (obviously increased from the value of 0.80 at the starting sample size) or the precision of the effect estimate (probability of obtaining a sample confidence interval with a width less than EW, given that EW fulfilled the precision needs of the study). In addition, during the iterative second step of the "two-step" procedure, it has been described the sawtooth behavior of the power, of the PW_{LE} function, and of the probability of obtaining a 95% CI width

© CIC Edizioni Internazionali

shorter than EW and a statistically significant result ($PW_{LE \cap SS}$). Furthermore, Cesana (19) showed that in order to combine the test power and the CI power, the "two step" procedure could also be implemented for a starting sample size calculated for a required CI width (sensibly chosen according to H_A), but considering also the test power of the increased sample size, as an useful extension of Beal's proposal (12) based only on the CI power.

The aim of this paper is to propose a new sample size calculation method in agreement with the regulatory requirements and the first priority of a researcher of demonstrating a clinically relevant difference; this method is naturally based on the intrinsic relationship between the power-based sample size and the precision of the effect estimate and goes beyond the limits of the usual power-based calculation approach that does not consider the precision of the effect estimation. This method is based on: (i) calculating the power-based sample size; (ii) considering the precision of the effect estimate by calculating the standardized Expected Half-width (EH) of the CI together with the probability of the event W_E given by sample CI standardized half-widths less than EH, conditional on the coverage (the interval contains the parameter under the alternative hypothesis), say $P(W_E|C)$; furthermore, a very broad scenario of the precision of the effect estimation given by the power-based sample size can be shown by calculating several values of standardized half-widths (H_j) together with the conditional probability of the event W_{Hj} given by sample standardized CI half-widths less than H_j, say $P(W_{Hj}|C)$; (iii) thereafter, iteratively increasing the starting power-based sample size to achieve an appropriate value of the joint probability function, conditional on the coverage $P(S \cap W_E|C)$, that combines the power of the statistical test (say, P(S) as the probability of the event S, given by a statistically significant result) and the power of the confidence interval $P(W_E)$; this step it is shown by considering EH as an appropriate and meaningful precision threshold, but it can be easily adapted to any chosen precision threshold. Nevertheless, it is important to focus that EH together with $P(W_E|C)$ can be considered as a relevant starting point for taking into account the precision of a research, given the power-based sample size.

Furthermore, steps (i), and (ii) can be carried out at different values of the test power (say, 0.80, 0.85 and 0.90) for achieving the best balance between the needs of the effect precision estimate together with the power of the study and the actual feasibility of the study.

The proposed method considers the CI coverage according to Lehmann's statement (20) that "there is no merit in short intervals that are far away from the true". However, it can be argued at least from a pragmatic point of view, that in sample size calculation the parameter under H_A is not the fixed unknown parameter representing the true state of nature (16), but is just the minimal clinically relevant difference (the lowest threshold of the true state of nature) that needs to be chosen for calculating conservatively the greatest sample size; in addition, as a further pragmatic consideration, the sample sizes obtained conditionally or unconditionally on the coverage are practically the same for the most frequently considered alternative hypotheses in the biomedical research (12), even in the case of a moderate heteroskedasticity (14).

The proposed method will be described for the case of one mean (or the difference between one mean and a reference value or the difference of two paired means) and the difference of two means of Gaussian distributed variables using one- or two-sample Student's t-tests as particular cases of scalar hypotheses in linear models.

STATISTICAL METHODOLOGY

Power-Based Sample Size Calculation

In the case of the one-sample Student's t-test, the null hypothesis is H_0: $\mu = \mu_0$ (μ_0 being a reference value or 0 for the paired-data case as in the crossover design of controlled clinical trials) against an alternative hypothesis such as H_A: $\mu = \mu_A$, where $\mu_A - \mu_0 = \delta$. In the case of the two-sample Student's t-test as in a controlled clinical trial according to a parallel group design, the null hypothesis is H_0: $\mu_1 - \mu_2 = 0$ against H_A: $\mu_1 - \mu_2 = \delta$. So, δ is the clinically relevant difference.

It has to be stressed that even if the analysis of variance is the standard statistical approach for the two-treatment two-period crossover trials analy-

© CIC Edizioni Internazionali

sis, their sample size calculation is usually based on a paired t-test; furthermore the period adjusted t-test for the treatment effect gives the same probability result as the analysis of variance (21).

The power-based sample sizes (n_0) for demonstrating δ are obtained by iteratively increasing a starting value (n in one group or n in each group with equal numbers), and consequently the non-centrality parameter λ ($\delta/\sigma\sqrt{k}$ with k = 1/n for the one-sample test, or 2/n for the two-sample test) of a non-central *t* distribution, until the area from $-\infty$ to $t_{1-\alpha/2,\nu}$ of a non-central *t* distribution is lower than the probability of the type II error (β); $t_{1-\alpha/2,\nu}$ is the $(1-\alpha/2)$ quantile of a central *t* distribution with ν degrees of freedom (n_0-1 for the one-sample test or $2(n_0-1)$ for the two-sample test), and α is the significance level (two sided).

This step can be easily carried out, for example, by means of the SAS® (22) functions Tinv and Probt: the first gives the quantiles of central or non central *t* distributions and the latter the cumulative density function of central or non central *t* distributions. Furthermore, this sample size calculation can be carried out by resolving by numerical method the integral of the product of a standardized Gaussian distribution and a central χ^2 distribution according to equation 1.1 shown in Appendix 1.

Confidence Interval Standardized Expected Half-Width (EH), Unconditional Confidence Interval Power ($P(W_E)$) and Conditional Confidence Interval Power $P(W_E|C)$.

Given the power-based sample size, the CI half-width under H_A is calculated as $t_{1-\alpha/2,\nu}\sigma\sqrt{k}$, since σ is assumed to be known at the sample size calculation phase; the standardized half-width ($t_{1-\alpha/2,\nu}\sigma\sqrt{k}$) is the standardized Expected Half-width (EH) (10, 23). Indeed, the expected value of the sample standardized half-width of the confidence interval $t_{1-\alpha/2,\nu}\sigma\sqrt{k}/\sigma$) is $t_{1-\alpha/2,\nu}\sigma\sqrt{k}$, since $\sigma\sqrt{\nu}/\sigma$ is a chi-distributed variable (χ_ν) with n degrees of freedom. In fact, we have:

$$E[\chi_\nu] = E\left[\frac{\sqrt{\nu}s}{\sigma}\right] = \sqrt{2}\frac{\Gamma[(\nu+1)/2]}{\Gamma(\nu/2)} \rightarrow$$

$$E[s] = \sqrt{\frac{2}{\nu}}\frac{\Gamma[(\nu+1)/2]}{\Gamma(\nu/2)}\sigma \approx \sigma$$

with a negligible bias (<2.8%) for n > 10 (24).

The unconditional confidence interval power, considered as the probability of obtaining sample standardized CI half-widths ≤ EH, is defined as: $P(W_E) = \Pr\{t_{1-\alpha/2,\nu}\sigma\sqrt{k}\, s \le t_{1-\alpha/2,\nu}\sigma\sqrt{k}\sigma\}$ and it can be calculated by resorting to the χ^2 distribution with n degrees of freedom, since $\nu s^2/\sigma^2$ is a χ^2_ν distributed variable with ν degrees of freedom. Since

$$P(W_E) = P\left(\chi^2_\nu \le \frac{\nu EH^2}{t^2_{1-\alpha/2,\nu}k}\right)$$

$P(W_E)$ can be obtained by integrating a χ^2 density function from 0 to $\nu EH^2/(t_{1-\alpha/2,\nu}\sigma\sqrt{k})^2$ equal to ν, being $EH = t_{1-\alpha/2,\nu}\sigma\sqrt{k}$.

This calculation, that with $EH^2 = h^2/\sigma^2$ corresponds to equation (1) of Grieve (13), can be easily carried out, for example, by means of the SAS® (22) function Probchi that gives the cumulative density function of central or non central χ^2 distributions.

Furthermore, the CI Power conditional on the coverage ($P(W_E|C)$) is obtained by dividing the joint probability that the standardized sample confidence interval is ≤ than EH and includes the parameter under H_A [$P(W_E \cap C)$] by the confidence coefficient ($1-\alpha$) or, equivalently, by the probability coverage P(C); $P(W_E \cap C)$ is calculated, owing to their independence, by the integral of the product of a central χ^2 distribution [$f_{x^2}(x^2)$] and a standardized Gaussian distribution [($f_Z(z)$], according to equation 2.4 in Appendix 2, as:

$$P(W_E \cap C) = \int_{-c\sqrt{x_E}}^{c\sqrt{x_E}} \left\{F_{X^2}(x_E) - F_{X^2}\left[(z/c)^2\right]\right\} f_Z(z)\,dz = 2\int_0^{c\sqrt{x_E}} \left\{F_{X^2}(x_E) - F_{X^2}\left[(z/c)^2\right]\right\} f_Z(z)\,dz$$

In addition to $P(W_E|C)$, given by $P(W_E \cap C)/(1-\alpha)$, this equation allows calculating relevant conditional probabilities ($P(W_{Hj}|C)$) together with other (standardized) CI half-widths (H_j) at which sample standardized CI half-widths (W_x) are expected to be lower (see Appendix 2); these values give an exhaustive scenario of the precision of the effect estimate allowable by the power-based sample size together with the probability of obtaining it and, furthermore, they can suggest a sensible choice of the precision threshold instead of EH. Finally, if the researcher's precision choice is sensibly expressed in the terms of a percentage of the effect size ($H'_j = p\delta/100\sigma$), this equation can be

© CIC Edizioni Internazionali

used for calculating $P(W_{H'j} \cap C)$ and then $P(W_{H'j}|C)$.

Conditional Confidence Interval Power and Statistical Test Power $P(S \cap W_E|C)$.

Furthermore, by including in equation 2.4 (Appendix 2) the condition of obtaining a statistically significant Student's t-test result (see Appendix 1), it is possible to obtain the joint probability of a statistically significant result (S), of a standardized sample CI half-width ≤ than EH (W_E), and of coverage (C), say $P(S \cap W_E \cap C)$, that, divided by the confidence coefficient (1 - α), gives $P(S \cap W_E|C)$. $P(S \cap W_E \cap C)$ is obtained, owing to their independence, by the product of a central χ^2 distribution [$f_{x^2}(x^2)$] and a standardized Gaussian distribution [$f_Z(z)$] according to equation 3.1 shown in Appendix 3:

$$P(S \cap W_E \cap C) = \int_{-\lambda/2}^{-\lambda+c\cdot\sqrt{x_E}} \left[F_{X^2}\left(\frac{z+\lambda}{c}\right)^2 - F_{X^2}\left(\frac{z}{c}\right)^2 \right] f_Z(z)\,dz + \int_{-\lambda+c\cdot\sqrt{x_E}}^{c\cdot\sqrt{x_E}} \left[F_{X^2}(x_E) - F_{X^2}\left(\frac{z}{c}\right)^2 \right] f_Z(z)\,dz$$

Sample Size calculation to obtain satisfactory values of $P(S \cap W_E|C)$.

Assuming that EH, given a power-based sample size for a test power ≥ 0.80, is satisfactory enough for the precision needs of the study (otherwise greater power values of the statistical test allow obtaining increasingly narrower EH values), the starting power-based sample size (n_0) is iteratively increased until $P(S \cap W_E|C)$ reaches a satisfactory enough value. It has to stress that as an alternative precision threshold the researcher can choose an H_j value or also a percentage of the effect size.

Defining n_1 a new value of the increased sample size, it is possible to calculate the corresponding degrees of freedom ν_1, the increased λ_1 ($\delta/\sigma\sqrt{k_1}$ with $k_1 = 1/n_1$ for the one-sample test, or $2/n_1$ for the two-sample test), and the (1-α/2) quantile of a central *t* distribution with n_1 degrees of freedom ($t_{1-\alpha/2,\nu 1}$) at the two-sided significance level (α). Then, the increased values of $P(S \cap W_E \cap C)$ can be obtained by equation 3.1 (Appendix 3) after having appropriately changed the upper and lower integration limits as follows:

$$P(S \cap W_E \cap C) = \int_{-\lambda_1/2}^{-\lambda_1+c_1\cdot\sqrt{x_1}} \left[F_X\left(\frac{z+\lambda_1}{c_1}\right)^2 - F_X\left(\frac{z}{c_1}\right)^2 \right] f_Z(z)\,dz + \int_{-\lambda_1+c_1\cdot\sqrt{x_1}}^{c_1\cdot\sqrt{x_1}} \left\{ F_X(x_1) - F_X\left(\frac{z}{c_1}\right)^2 \right\} f_Z(z)\,dz$$

where: $c_1 = t_{1-\alpha/2,\nu 1} / \sqrt{\nu_1}$, and $x_1 = (t_{1-\alpha/2,\nu 0} / t_{1-\alpha/2,\nu 1})^2 (n_1 \cdot \nu_1 / n_0)$, being $t_{1-\alpha/2,\nu 0}$ the (1-α/2) quantile of a central *t* distribution with ν_0 degrees of freedom from the starting power-based sample size (n_0). Thereafter, $P(S \cap W_E|C)$ is obtained by dividing $P(S \cap W_E \cap C)$ by the confidence coefficient (1-α).

RESULTS

The power-based sample sizes (n_0) for demonstrating an effect size (δ/σ) of 0.10(0.025) 0.20(0.05) 0.7(0.1) 1.3 at a test power (1-β) of ≥0.80 and significance level (α) of 0.05 (two-sided) have been calculated for the one-sample and the two-sample Student's t-test.

For the one-sample t-test, Table 1 shows the effect size in the first column, the power-based sample size (n_0) in the second, the standardized expected width (EH) in the third and the values of $P(W_E|C)$ and of $P(S \cap W_E|C)$ given n_0 in the fourth and fifth column, respectively. The next three columns show precision thresholds (H_j) under which sample standardized CI half-widths are expected to be less at $P(W_{Hj}|C)$ values of 0.80, 0.90, and 0.99. Finally the last five columns show the sample sizes necessary for obtaining $P(S \cap W_E|C)$ values at least of 0.70, 0.75, 0.80, 0.85, and 0.90. With the following example of the proposed sample size calculation approach, the expected values are given together with the results (between brackets and in italic) from a simulation study on 10,000 of one sample. The parameters of the simulation study have been chosen accordingly to the effect size of 0.6, and the power-based calculated sample size is of 24 patients for a test power of 0.80 and a significance level of 0.05 (two-sided).

Hermansen et al. (25) compared the acute bronchodilatory effect of the long-acting β_2-agonist formoterol against the short-acting β_2-agonist terbutaline during exercise-induced bronchoconstriction in children with persistent asthma according to a cross-over design; they enrolled in the

© CIC Edizioni Internazionali

Table 1. One sample: power-based sample sizes ($1-\beta \geq 0.80$ and $\alpha = 0.05$, two-sided) for some effect size values (δ/σ), together with their corresponding standardized half-width (EH), $P(W_E|C)$ and $P(S \cap W_E|C)$; three thresholds (H_j) of the half-width CI under which the sample half-widths are expected to be lower at conditional probabilities $[P(W_{Hj}|C)]$ of 0.80, 0.90 and 0.99 and five sample sizes for $[P(S \cap W_E|C)]$ of 0.70, 0.75, 0.80, 0.85, and 0.90.

δ/σ	n_0	EH	$P(W_E\|C)$	$P(S \cap W_E\|C)$	$P(W_{Hj}\|C)$			$P(S \cap W_E\|C)$				
					0.80	0.90	0.99	0.70	0.75	0.80	0.85	0.90
0.100	787	0.0699	0.50429	0.41802	0.0714	0.0722	0.0741	825	835	849	877	985
0.125	505	0.0874	0.50536	0.42078	0.0897	0.0909	0.0938	535	542	553	571	631
0.150	351	0.1049	0.50643	0.42315	0.1083	0.1100	0.1143	375	381	389	402	440
0.175	259	0.1224	0.50750	0.42612	0.1268	0.1292	0.1349	279	284	291	301	324
0.200	199	0.1398	0.50857	0.42887	0.1456	0.1487	0.1562	216	221	226	234	250
0.250	128	0.1749	0.51073	0.43381	0.1839	0.1888	0.2007	141	145	149	154	164
0.300	90	0.2095	0.51287	0.43989	0.2222	0.2293	0.2464	101	103	106	110	117
0.350	67	0.2439	0.51500	0.44584	0.2611	0.2708	0.2939	76	78	80	84	88
0.400	52	0.2784	0.51714	0.45160	0.3007	0.3132	0.3434	60	61	63	66	69
0.450	41	0.3156	0.51946	0.45442	0.3439	0.3600	0.3989	48	49	51	53	56
0.500	34	0.3489	0.52154	0.46134	0.3831	0.4028	0.4504	40	41	43	44	47
0.550	28	0.3878	0.52397	0.46347	0.4296	0.4536	0.5124	33	34	36	37	39
0.600	24	0.4223	0.52613	0.46919	0.4712	0.5000	0.5695	29	30	31	32	34
0.650	21	0.4552	0.52819	0.47560	0.5114	0.5448	0.6253	25	26	27	28	30
0.700	19	0.4819	0.52987	0.48475	0.5448	0.5814	0.6720	23	24	25	26	27
0.800	15	0.5538	0.53437	0.49240	0.6340	0.6822	0.8018	18	19	20	21	22
0.900	12	0.6354	0.53946	0.49670	0.7382	0.8002	0.9559	15	15	16	17	18
1.000	10	0.7154	0.54437	0.50243	0.8413	0.9188	1.1146	12	13	14	14	15
1.100	9	0.7687	0.54760	0.51466	0.9115	1.0003	1.2224	11	12	12	13	14
1.200	8	0.8360	0.55160	0.52231	0.9992	1.1036	1.3646	10	10	11	11	12
1.300	7	0.9248	0.55673	0.52621	1.1168	1.2108	1.5544	9	9	10	10	11

study 24 patients since "this sample size, based on their experience from previous studies, would give a 80% chance to detect a difference of 8% (corresponding to approximately 0.2 L) in the Forced Expiratory Volume in one second (FEV_1) at a two-sided test at a 5% significance level". This difference is consistent with an effect size of about 0.6, considering the variability of the FEV_1 of about 0.33 shown in their Tables 1 and 2 and from their references. In order to plan a cross-over clinical trial for comparing the effect of a new bronchodilator against the standard drug, the power-based calculated sample size based on these estimates (effect size: $\delta/\sigma = 0.6$) is 24 for a power of 0.80 at a one-sample Student's t-test with $\alpha = 0.05$ (two-sided). Therefore, EH is 0.4223, $P(W_E|C)$ is 0.5261 *(0.5083)* and $P(S \cap W_E|C)$ is 0.4691 *(0.4323)*.

Table 1 shows that a sample size of 34 gives a $P(S \cap W_E|C)$ of 0.90 *(0.9044)*; however, a sample size of 31 can also be a sensible choice since $P(S \cap W_E|C)$ is at least 0.80 *(0.8287)*.

In addition, it is possible to see that a $P(W_{Hj}|C)$ value of 0.80 *(0.7979)* is obtained if a precision threshold of 0.4712 is chosen instead of the EH (equal to 0.4223); furthermore, values of 0.90 or 0.99 are given by the thresholds of 0.5000 *(0.9028)* or of 0.5695 *(0.9899)*, respectively. More interestingly, selecting one of these thresholds, a $P(S \cap W_{Hj}|C)$ value of 0.80 is obtained with n = 27 for 0.4712 *(0.8074)*, n = 26 for 0.5000 *(0.8320)* and, finally, with just the starting sample size of 24 for 0.5695

© CIC Edizioni Internazionali

Table 2. Two samples: power-based sample sizes ($1-\beta \geq 0.80$ and $\alpha = 0.05$, two-sided) for some effect size values (δ/σ), together with their corresponding standardized half-width (EH), $P(W_E|C)$ and $P(S \cap W_E|C)$; three thresholds (H_j) of the half-width CI under which the sample half-widths are expected to be lower at conditional probabilities [$P(W_{Hj}|C)$] of 0.80, 0.90 and 0.99 and five sample sizes for [$P(S \cap W_E|C)$] of 0.70, 0.75, 0.80, 0.85, and 0.90.

δ/σ	n_0	EH	$P(W_E\|C)$	$P(S \cap W_E\|C)$	$P(W_{Hj}\|C)$			$P(S \cap W_E\|C$				
					0.80	0.90	0.99	0.70	0.75	0.80	0.85	0.90
0.100	1571	0.0699	0.50214	0.41297	0.0706	0.0710	0.0720	1611	1623	1640	1712	1967
0.125	1006	0.0874	0.50267	0.41427	0.0885	0.0892	0.0906	1038	1047	1060	1099	1259
0.150	699	0.1049	0.50321	0.41556	0.1065	0.1074	0.1095	725	732	743	768	857
0.175	514	0.1224	0.50375	0.41689	0.1246	0.1258	0.1287	536	542	550	569	643
0.200	394	0.1399	0.50428	0.41828	0.1427	0.1443	0.1481	413	418	425	439	493
0.250	253	0.1747	0.50535	0.42118	0.1792	0.1817	0.1875	268	272	277	286	316
0.300	176	0.2097	0.50643	0.42371	0.2162	0.2197	0.2282	188	191	195	202	220
0.350	130	0.2443	0.50750	0.42688	0.2531	0.2579	0.2694	140	143	146	151	163
0.400	100	0.2789	0.50857	0.42986	0.2904	0.2967	0.3117	109	111	114	118	126
0.450	79	0.3143	0.50967	0.43187	0.3289	0.3369	0.3560	87	89	91	94	100
0.500	64	0.3498	0.51078	0.43389	0.3679	0.3778	0.4016	71	73	75	77	82
0.550	53	0.3852	0.51188	0.43621	0.4070	0.4191	0.4480	59	61	62	65	69
0.600	45	0.4189	0.51294	0.44001	0.4448	0.4589	0.4932	51	52	53	55	59
0.650	39	0.4510	0.51395	0.44489	0.4808	0.4973	0.5371	44	45	47	48	51
0.700	34	0.4842	0.51500	0.44862	0.5184	0.5376	0.5835	39	40	41	42	45
0.800	26	0.5571	0.51732	0.45186	0.6020	0.6272	0.6886	30	31	32	33	35
0.900	21	0.6237	0.51946	0.45883	0.6796	0.7114	0.7883	25	25	26	27	29
1.000	17	0.6987	0.52189	0.46176	0.7683	0.8080	0.9050	20	21	22	22	24
1.100	15	0.7479	0.52350	0.47488	0.8272	0.8730	0.9844	18	18	19	20	21
1.200	12	0.8467	0.52677	0.46979	0.9474	1.0050	1.1487	15	15	16	16	17
1.300	11	0.8895	0.52819	0.48317	0.9994	1.0645	1.2219	13	14	14	15	16

(0.8130)(data not shown). Otherwise, if the researcher prefers to choose a precision threshold as a percentage of the effect size, as it seems sensible for having a more direct relationship between the precision and the effect size, it is possible to obtain (SAS program available from the author) that, for example, a $P(S \cap W_{H'j}|C)$ value of 0.80 is given by 31 subjects *(0.8082)* if the threshold is the 75% of the effect size (0.45) and by 26 subjects *(0.7999)* if the threshold is the 80% (0.48) (data not shown). The sample sizes needed to obtain a $P(S \cap W_E|C)$ of 0.80 are, obviously, more than those needed to obtain only a statistical test power of 0.80, but the increases are less than 10% for an effect size ≤ 0.125, less than 20% up to an effect size of 0.350, less than 30% up to an effect size of 0.650, and less than 40% for an effect size up to 1.3 for which the increase is only of a few subjects.

Table 2 shows the results of the proposed method for the two-sample Student's t-test; the sample sizes are for each group.

With the following example of the proposed sample size calculation approach the expected values are given together with the results (between brackets and in italic) from a simulation study on 10,000 trials with two samples. The parameters of the simulation study have been chosen accordingly to an effect size of 0.5 (particularly, $\mu_1 = 6.0$, $\mu_2 = 8.5$, $\sigma_1 = \sigma_2 = 5$), and the power-based calculated sample size is of 64 patients in each treatment group for a test power of 0.80 and a significance level of 0.05 (two-sided).

© CIC Edizioni Internazionali

Kraus et al. (26) reported the results of a parallel controlled clinical trial (STRRIDE) aimed to investigate the effects of the amount and intensity of the physical exercise on the serum lipoproteins profile in patients with mild-to-moderate dyslipidemia. Four groups were compared: control, low-amount and moderate intensity, low amount and high intensity, and, finally, high amount and high intensity exercise for a six months period. A statistical significant difference against the control group has been found only for the group of patients randomized to the high amount and high intensity exercise; particularly, this group showed a mean HDL cholesterol concentration increase of 4.3 mg/dl against a decrease of 0.6 mg/dl in the control one (Figure 2 on page 1489) with a difference of 6.5 mg/dl (48.6 – 42.1 mg/dl) between these two groups at the end of the six months study. Furthermore, from the standard error values shown in Table 2 (page 1487) for the HDL base and end of study, respectively, it is straightforward to obtain a value of about 13 mg/dl as an estimate of the phenomenon variability of the HDL serum concentration. Suppose that a researcher wants to confirm these results on a different kind of patients or in a different country (bridging study); so he/she designs a new study for comparing no physical exercise (control group) to a high amount and high intensity exercise based on these estimates, since the difference of the HDL cholesterol serum concentration can be sensibly chosen as the primary objective of the clinical trial. For δ = 6.5 mg/dl and σ = 13 mg/dl, giving an effect size (δ/σ) of 0.5, the power-based calculated sample size is 64 for having a power of at least 0.80 at a two-sample Student's t-test with α = 0.05 (two-sided). Therefore, EH is 0.3498, $P(W_E|C)$ is 0.5108 *(0.5083)* and $P(S\cap W_E|C)$ is 0.4339 *(0.4324)*.

Table 2 also shows that, with a 17.2% sample size increase (75 subjects in each group) $P(S\cap W_E|C)$ is at least 0.80 *(0.8254)*; a greater increase of 28.1% (82 subjects in each group) leads to the very high value of 0.90 *(0.9088)* of obtaining both of the relevant research aims given by the $P(S\cap W_E|C)$ function.

In addition, it is worthwhile to stress that a precision threshold of 0.3679 instead of the EH (equal to 0.3498) gives a $P(W_{Hj}|C)$ value of 0.80 *(0.7971)*; furthermore, thresholds of 0.3778 and 0.4016 give values of 0.90 *(0.8965)* and 0.99 *(0.9909)*, respectively. More interestingly, selecting one of these thresholds a $P(S\cap W_{Hj}|C)$ of at least 0.80 is obtained with n = 70 for 0.3679 *(0.8204)*, and with n = 67 for 0.3778 *(0.8049)* and, finally, with just the starting sample size of 64 for 0.4016 for which $P(S\cap W_{Hj}|C)$ is 0.8103 *(0.8096)* (data not shown). Otherwise, for a precision threshold in the terms of a percentage of the effect size of 75% equal to 0.3750 and of 80% equal to 0.4000, for example, a $P(S\cap W_{H'j}|C)$ value of 0.80 is obtained with 68 subjects *(0.8158)* and 64 *(0.8084)*, respectively (data not shown).

Of course, also in the case of the two-sample Student's t-test, the sample sizes necessary for a $P(S\cap W_E|C)$ of 0.80 are more than that needed for obtaining only a test power of 0.80, but less than 10% up to an effect size of 0.250, less than 20% up to an effect size of 0.70, and less than 30% for effect size values of up to 1.3 (at which the absolute increase is only of a few units).

Considering the sample sizes shown in Table 2 of Jiroutek et al. (16), a probability value of 0.80 (0.90) for a required standardized width of 0.5 and for demonstrating an effect size of 0.5, given validity, is obtained with 134 (138) subjects in each group, against 75 (82) subjects from the proposed procedure. So, in this case, there is a 28.6% precision reduction (about 0.7 against 0.5), but the sample size increase is only of 17.2% (28.1%) against the 109.4% (115.6%) shown by Jiroutek et al. (16).

DISCUSSION

The unsatisfactory aspects of the two distinct sample size calculation procedures for demonstrating and estimating an effect can be overcome if they are appropriately combined bearing in mind that the power-based approach is more important to researchers who are mainly interested in demonstrating a clinically relevant effect and, in addition, it is the recommended approach by the standardized guidelines (17) according to ethical principles.

It is worth underlining that the starting power-based sample size can provide researchers with

© CIC Edizioni Internazionali

very useful information that should be taken into account when planning a study. In particular: i) the achievable level of precision (EH); ii) the unconditional or the conditional probability of obtaining a standardized CI half-width of less than EH [$P(W_E)$ or $P(W_E|C)$] which is however always unsatisfactorily about 0.50 regardless of the chosen power or the study; iii) some other conditional probability [$P(W_{Hj}|C)$] values of obtaining other precision levels (H_j) which, more than EH (lower precision levels), can be obtained at probability values of more than $P(W_E|C)$; and iv) the most relevant probability of demonstrating a clinically relevant hypothesis and obtaining the precision level of the study $P(S\cap W_E|C)$, which is even less than the unsatisfactory $P(W_E|C)$.

Consequently, the starting power-based sample size can be increased for obtaining satisfactory values of $P(W_E|C)$ and, above all, of $P(S\cap W_E|C)$.

Particularly, starting from a test power of 0.80, the researcher can accept EH as a suitable precision threshold or select one of the H_j or a percentage of the effect size (H'_j); however, it has to stress that both have to be >EH since, otherwise, more subjects have to be enrolled into the trial only on the basis of the required precision of the effect estimate, not consistently with the regulatory and ethical requirements focused on the necessary sample size for demonstrating a clinical relevant difference. However, if EH is too higher for the precision needs of the study, a sample size calculated for a greater test power (0.85 or 0.90) allows obtaining lower EH or H_j or H'j as suitable values of the precision threshold.

The sample size calculation method for obtaining satisfactory $P(S\cap W_E|C)$ values concentrates on the main interest of researchers; this function plays a pivotal role insofar as it links statistical testing and estimating procedures by combining the two powers of inferential statistical procedures, and gives the probability of achieving the two aims of demonstrating a relevant difference and estimating it with satisfactory precision.

The results of the simulation study for the one sample and the two samples case are very near to their expected values (lower than 1%); however, in some cases, a difference of about 2-3% is due to the fact that the Tables show values of $P(S\cap W_E|C)$ and $P(W_{Hj}|C)$ at least equal to selected probability values, and, finally, it has to consider that for the shown examples at each one or two unities increase of the sample size $P(S\cap W_E|C)$ (and also $P(S\cap W_{Hj}|C)$) increases of 0.05. Particularly, from the Table 1 it can be seen that the expected value of $P(S\cap W_E|C)$ is at least 0.75 for $n_1 = 30$, at least 0.80 for $n_1 = 31$, and at least 0.85 for $n_1 = 32$; so, for example, it is absolutely acceptable the result of 0.8287 instead of at least 0.80 from the simulation study with $n_1 = 31$.

It is also worth considering that this procedure can be used even if researchers do not have a previous appropriate estimate of the variability of the phenomenon, and therefore they have to rely on the effect size.

The increase in the usual power-based sample size required for a $P(S\cap W_E|C)$ value of 0.80 is generally acceptable and counterbalanced by the greater validity of the planned research.

SAS® sample size calculation programs are available from the author.

APPENDIX 1

For a unilateral alternative hypothesis $H_A: \mu_A - \mu_0 = \delta$ or $H_A: \mu_1 - \mu_2 = \delta$ with $\delta > 0$, the event rejection of the null hypothesis [S given by $t'_\nu \geq t_{1-\alpha/2,\nu}$, being t'_ν the sample test statistic, and $t_{1-\alpha/2,\nu}$ the 1-$\alpha/2$ quantile of a central t distribution with ν degrees of freedom equal to $r(n_0-1)$ with r = 1 or 2 for the one or the two samples test, respectively] is obtained at a power value given by $P(S) = P[t'_\nu(\lambda) \geq t_{1-\alpha/2,\nu}] \geq 1-\beta$ where $t'_\nu(\lambda)$ is the quantile of a non-central *t* distribution with ν degrees of freedom and non-centrality parameter $\lambda = \delta/\sigma\sqrt{k}$ with $k = r/n_0$, being n_0 the sample size in the one sample or in each sample. Since $t'_\nu(\lambda) = (Z+l)/(Z\sqrt{\nu})$, being Z and X independent random variables distributed as a standard normal [N(0,1)] and a χ with n degrees of freedom, respectively, we have:

$$P\left\{(Z+\lambda)/\left(X/\sqrt{\nu}\right) \geq t_{1-\alpha/2,\nu}\right\} =$$
$$= P\left\{Z+\lambda \geq t_{1-\alpha/2,\nu}\, X/\sqrt{\nu}\right\} = P\left\{Z \geq -\lambda + cX\right\}$$

putting $c = t_{1-\alpha/2,\nu}/\sqrt{\nu}$, leads, according to Johnson et al (27), to calculate the test power as:

© CIC Edizioni Internazionali

$$P(S) = \int_0^{\infty} \int_{-\lambda+cx}^{\infty} f_X(x) f_Z(z) dx\, dz$$

where $f_X(x)$ is a χ density function and $f_Z(z)$ is a standard normal density function.
By changing from the χ variable (X) to a χ^2 variable (X^2), and, accordingly, the integration limits, integrating before for x^2 and after for z, and then putting $F_{x^2}(x^2) = \int_0^{x^2} f_{x^2}(t)$ as the cumulative central χ^2 distribution function, we obtain:

$$P(S) = \int_{-\lambda}^{\infty} \left(\int_0^{((z+\lambda)/c)^2} f_{X^2}(t)\, dt \right) f_Z(z)\, dz = \\ = \int_{-\lambda}^{\infty} F_{X^2}\left[\left((z+\lambda)/c \right)^2 \right] f_Z(z)\, dz \quad (1.1)$$

The inner integral can be easily obtained by a statistical software (22: function Probchi, for example) and, since the standard normal distribution remains finite over its range, the double integral can be resolved by numerical integration by using, for example, the QUAD function of SAS®/IML (28) without problems of overflow in the calculation. The test power function reaches the required value ($1-\beta = 0.80$, at least) at a sample size value, say n_0, giving $k_0 = r/n_0$ and $\nu_0 = r(n_0 - 1)$.

APPENDIX 2

Let's define W_E the event that the standardized sample half-width (W_x) is $\leq$ EH; since $W_z = t_{1-\alpha/2,\nu} s\sqrt{k}/\sigma$ and $EH = t_{1-\alpha/2,\nu}\sqrt{k}$, we have that:

$$W_E = \{W_x \leq EH\} = \{t_{1-\alpha/2,\nu}\, s\sqrt{k}/\sigma \leq EH\} = \\ = \{s/\sigma \leq EH/(t_{1-\alpha/2,\nu}\sqrt{k})\}$$

then, by multiplying both sides by $\sqrt{\nu}$, we obtain:

$$W_E = \left\{ \frac{\sqrt{\nu}\, s}{\sigma} \leq \frac{\sqrt{\nu}\, EH}{t_{1-\alpha/2,\nu}\sqrt{k}} \right\} = \left\{ X \leq \sqrt{x_E} \right\} \quad (2.1)$$

where $X = \sqrt{\nu}$, s/σ is a chi distributed variable with ν degrees of freedom (χ_ν) and $x_E = (\nu\, EH^2)/(t_{1-\alpha/2,\nu}\sqrt{k})^2$ is, obviously, equal to ν.
By defining the event coverage that the CI includes the parameter δ, estimated by $\hat{\delta}$, as:

$C = \left\{ \hat{\delta} - t_{1-\alpha/2,\nu}\, s\sqrt{k} \leq \delta \leq \hat{\delta} + t_{1-\alpha/2,\nu}\, s\sqrt{k} \right\}$, by rearranging and dividing for $\sigma\sqrt{k}$, we have:

$$C = \left\{ -\frac{t_{1-\alpha/2,\nu}\, s\sqrt{k}}{\sigma\sqrt{k}} \leq \frac{\hat{\delta}-\delta}{\sigma\sqrt{k}} \leq \frac{t_{1-\alpha/2,\nu}\, s\sqrt{k}}{\sigma\sqrt{k}} \right\} = \\ = \left\{ -\frac{t_{1-\alpha/2,\nu}\, s}{\sigma} \leq \frac{\hat{\delta}-\delta}{\sigma\sqrt{k}} \leq \frac{t_{1-\alpha/2,\nu}\, s}{\sigma} \right\}$$

Since $Z = \frac{\hat{\delta}-\delta}{\sigma\sqrt{k}}$,,
by multiplying the numerator and denominator of the two extreme terms by $\sqrt{\nu}$, this expression becomes:

$$C = \left\{ -\frac{t_{1-\alpha/2,\nu}\, s}{\sigma} \leq Z \leq \frac{t_{1-\alpha/2,\nu}\, s}{\sigma} \right\} = \\ = \left\{ -\frac{t_{1-\alpha/2,\nu}}{\sqrt{\nu}} \cdot \frac{\sqrt{\nu}\, s}{\sigma} \leq Z \leq \frac{t_{1-\alpha/2,\nu}}{\sqrt{\nu}} \cdot \frac{\sqrt{\nu}\, s}{\sigma} \right\} = \\ = \{-c\, X \leq Z \leq c\, X\} \quad (2.2)$$

$X = \sqrt{\nu}$, s/σ is a chi distributed variable with ν degrees of freedom (χ_ν) and $c = t_{1-a/2,\nu}/\sqrt{\nu}$.
Thereafter, the joint event $W_E \cap C$, expressed in the terms of a chi (X) and a standard normal (Z) distributed variables, is:

$$W_E \cap C = [(0 \leq X \leq \sqrt{x_E}) \cap (-c\, X \leq Z \leq c\, X)]$$

or, equivalently,

$$W_E \cap C = \left\{ \left(\frac{z}{c} \leq X \leq \sqrt{x_E} \right) \cap \left(-c\sqrt{x_E} \leq Z \leq c\sqrt{x_E} \right) \right\}$$

So, by changing from a chi variable (X) to a chi-squared variable (X^2), its probability is given by:

$$P(W_E \cap C) = \int_{(z/c)^2}^{x_E} \int_{-c\sqrt{x_E}}^{c\sqrt{x_E}} f_{X^2}(t) f_Z(z)\, dt\, dz \quad (2.3)$$

where $f_{X^2}(t)$ is a chi-squared density function and $f_Z(z)$ is a standard normal density function.
By solving the double integral for x^2 before, and then for z, and putting $F_{x^2}(x^2) = \int_0^{x^2} f_{x^2}(t)\, dt$ as the cumulative central χ^2 distribution function, we obtain, according to equation (2) of Grieve (13), that:

$$P(W_E \cap C) = \int_{-c\sqrt{x_E}}^{c\sqrt{x_E}} \left\{ F_{X^2}(x_E) - F_{X^2}\left[(z/c)^2 \right] \right\} f_Z(z)\, dz = \\ = 2\int_0^{c\sqrt{x_E}} \left\{ F_{X^2}(x_E) - F_{X^2}\left[(z/c)^2 \right] \right\} f_Z(z)\, dz \quad (2.4)$$

since $c = t_{1-\alpha/2,\nu}/\sqrt{\nu}$, the upper integration limit is equal to $t_{1-\alpha/2,\nu}$. The inner integral can be easily obtained by a statistical software (the function Probchi of SAS® (22), for example) and, since the standard normal distribution remains finite over its range, the double integral can be resolved by numerical integration (QUAD function of SAS®/IML (28), for example) without problems

© CIC Edizioni Internazionali

of overflow in the calculation. Finally, since P(C) = 1- a, from equation (2.4) it is straightforward to obtain $P(W_E|C) = P(W_E \cap C) / P(C)$ or $P(W_E \cap C) / (1- \alpha)$.

Equation 2.4 allows calculating relevant values of the unconditional probability [$P(W_{Hj} \cap C)$, say J = 0.70, 0.75,…, 0.95, 0.99] and then of the conditional one $P(W_H|C)$ by increasing the upper integration limit (say, $x_H > x_E$); thereafter the corresponding precision thresholds ($H_j > EH$) can be calculated as: $H_j = t_{1-\alpha/2,\nu} \sqrt{k\, x_H} / \sqrt{\nu}$.

Furthermore, equation 2.4 allows calculating $P(W_{H'j} \cap C)$ and then $P(W_{H'j}|C)$ by putting the $x_H = (\nu\, H'^2)/(t_{1-\alpha/2,\nu} \sqrt{k})^2$, being $H'j = p\delta/100\sigma$.

APPENDIX 3

The joint event $S \cap W_E \cap C$, expressed in the terms of Z and X, being Z and X independent random variables distributed as a standard normal [N(0,1)] and a chi with n degrees of freedom, is:

$$S \cap W_E \cap C = \left\{ (Z > -\lambda + c\,X) \cap \left(0 \le X \le \sqrt{x_E}\right) \cap (-c\,X \le Z \le c\,X) \right\}$$

or, equivalently:

$$S \cap W_E \cap C = \left\{ \left(0 \le X \le \sqrt{x_E}\right) \cap (Z > -\lambda + c\,X) \cap (-c\,X \le Z \le c\,X) \right\} = \\ = \left\{ \left(0 \le X \le \sqrt{x_E}\right) \cap \left(\max[-c\,X; -\lambda + c\,X] \le Z \le c\,X\right) \right\}$$

Considering the lower term of the second intersection: max[-cX; -λ+cX], -cX ≥ -λ +cX is obtained for X≤λ/2c; furthermore, since $X \le \sqrt{x_E}$, we have to take into account two cases of which the first is:

$$\sqrt{x_E} \ge \frac{\lambda}{2c} \rightarrow \sqrt{x_E} \ge \frac{\delta/\sigma\sqrt{k}}{2t_{1-\alpha/2,\nu}/\sqrt{x_E}} \rightarrow \\ \sqrt{x_E} \ge \frac{\delta\sqrt{x_E}}{2 \cdot t_{1-\alpha/2,\nu}\sigma\sqrt{k}} \rightarrow 2 \cdot t_{1-\alpha/2,\nu}\sqrt{k} \ge \frac{\delta}{\sigma} \rightarrow \frac{\delta}{\sigma} \le 2EH$$

since $t_{1-\alpha/2,\nu} \sqrt{k} = EH$.

So, in this first case with $\delta/\sigma \le 2EH$, max{-cX; -λ + cX} = -cX for $0 \le X \le \lambda / 2c$ and otherwise, for $\lambda/ 2c \le X \le \sqrt{x_E}$, max{-cX; -λ + cX} = -λ + cX.

The joint event $S \cap W_E \cap C$, expressed in the terms of X and Z, being X and Z independent random variables distributed as a chi with ν degrees of freedom and a standard normal [N(0,1)], is:

$$(S \cap W_E \cap C) = \left\{ \left[0 \le X \le \left(\frac{\lambda}{2c}\right)\right] \cap (-c\,X \le Z \le c\,X) \right\} \cup \\ \cup \left\{ \left[\left(\frac{\lambda}{2c}\right) \le X \le \sqrt{x_E}\right] \cap (-\lambda + c\,X \le Z \le c\,X) \right\}$$

Then, by solving the double integral for x^2 and then for z, and putting $F_{X^2}(x^2) = \int_0^{x^2} f_{x^2}(t)\,dt$ as the cumulative central χ^2 distribution function, we obtain:

$$P(S \cap W_E \cap C) = \int_{-\lambda/2}^{-\lambda + c\cdot\sqrt{x_E}} \left(\int_{(z/c)^2}^{((z+\lambda)/c)^2} f_{X^2}(t)\,dt \right) f_Z(z)\,dz + \\ + \int_{-\lambda + c\cdot\sqrt{x_E}}^{c\cdot\sqrt{x_E}} \left(\int_{(z/c)^2}^{x_E} f_{X^2}(t)\,dt \right) f_Z(z)\,dz = \\ = \int_{-\lambda/2}^{-\lambda + c\cdot\sqrt{x_E}} \left[F_{X^2}\left(\frac{z+\lambda}{c}\right)^2 - F_{X^2}\left(\frac{z}{c}\right)^2 \right] f_Z(z)\,dz + \\ + \int_{-\lambda + c\cdot\sqrt{x_E}}^{c\cdot\sqrt{x_E}} \left[F_{X^2}(x_E) - F_{X^2}\left(\frac{z}{c}\right)^2 \right] f_Z(z)\,dz \qquad (3.1)$$

where the inner integral can be easily obtained, for example, by the SAS® (22) function Probchi; the double integral can be resolved by numerical integration by QUAD function of SAS®/IML (28) without problems of overflow in the calculation.

Since $c = t_{1-\alpha/2,\nu}/\sqrt{\nu}$, $c\sqrt{x_E}$ is equal to $t_{1-\alpha/2,\nu}$.

The second case, that, however, does not occur in this paper since δ/σ is always < 2EH, is given by:

$$\sqrt{x_E} < \frac{\lambda}{2c} \rightarrow \sqrt{x_E} < \frac{\delta/\sigma\sqrt{k}}{2t_{1-\alpha/2;\nu}/\sqrt{x_E}} \rightarrow \\ \sqrt{x_E} < \frac{\delta\sqrt{x_E}}{2t_{1-\alpha/2;\nu}\sigma\sqrt{k}} \rightarrow 2t_{1-\alpha/2;\nu}\sqrt{k} < \frac{\delta}{\sigma} \rightarrow \frac{\delta}{\sigma} > 2EH$$

since $t_{1-\alpha/2,\nu} \sqrt{k} = EH$.

So, max{-cX; -λ + cX} = -cX for $0 \le X \le \sqrt{x_E}$.

The joint event $S \cap W_E \cap C$, expressed in the terms of X and Z, being X and Z independent random variables distributed as a chi with ν degrees of freedom and a standard normal [N(0,1)], is:

$$(S \cap W_E \cap C) = \left\{ \left[0 \le X \le \sqrt{x_E}\right] \cap (-\lambda + c\,X \le Z \le c\,X) \right\}$$

equal to $(W_E \cap C)$, and its probability is obtained by equation 2.4 in Appendix 2.

Equation 3.1 (and 2.4) allows obtaining $P(S \cap W_E|C)$ value at increasing sample size, appropriately changing the integration limits.

REFERENCES

1. Cohen J. Statistical Power Analysis for the Behavioral

© CIC Edizioni Internazionali

Sciences, New York: Academic Press, 1969.
2. McHugh RB, Le CT. Confidence Estimation and the Size of a Clinical Trial. Controlled Clinical Trials 1984; 5: 157-163.
3. Greenland S. On sample-size and power calculations for studies using confidence intervals. American Journal of Epidemiology 1988; 128: 231-237.
4. Bristol DR. Sample sizes for constructing confidence intervals and testing hypotheses. Statistics in Medicine 1989; 8: 803-811.
5. Goodman SN, Berlin JA. The Use of Predicted Confidence Intervals When Planning Experiments and the Misuse of Power When Interpreting Results. Annals of Internal Medicine 1994; 121: 200-206.
6. Borenstein M. Planning for precision in survival studies. Journal of Clinical Epidemiology 1994; 47: 1277-1285.
7. Nicewander WA, Price JM. A Consonance Criterion for Choosing Sample Size. The American Statistician 1997; 51: 311-317.
8. Feinstein AR, Concato J. The Quest for Power: Contradictory Hypotheses and Inflated Sample Sizes. Journal of Clinical Epidemiology 1998; 51: 537-545.
9. Hahn GJ, Meeker WQ. Statistical Intervals: a guide for practitioners, New York: John Wiley & Sons Inc., 1991.
10. Kupper LL, Hafner KB. How Appropriate Are Popular Sample Size Formulas?. The American Statistician 1989; 43: 101-105.
11. Grieve AP. Letter in reply to: Clinical Trial Numbers and Confidence Intervals of Pre-Specified Size by Day S.J, Lancet 1988, December 17 1989, 1427, Lancet 337.
12. Beal SL. Sample Size Determination for Confidence Intervals on the Population Mean and the Difference Between Two Population Means. Biometrics 1989; 45: 969-977.
13. Grieve AP. (1991). Confidence Intervals and Sample Sizes. Biometrics 1990; 47: 1597-1603.
14. Wang Y, Kupper LL. Optimal Sample Sizes for Estimating the Difference in Means between Two Normal Populations Treating Confidence Interval Length as a Random Variable. Communications in Statistics: Theory and Methods 1997; 26: 727-741.
15. Machin D, Campbell MJ, Fayers PM, Pinol APY. Sample Size Tables for Clinical Studies (2nd ed). Oxford U.K.: Blackwell Scientific Publications, 1997.
16. Jiroutek MR, Muller KE, Kupper LL, Stewart PW. A New Method for Choosing Sample Size for Confidence Interval-Based Inferences. Biometrics 2003; 50: 580-590.
17. ICH E9 Guideline 1998. Statistical Principles for Clinical Trials, (accessed October 2010). http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500002928.pdf
18. Cesana BM, Reina G, Marubini E. Sample Size for Testing a Proportion in Clinical Trials: a "Two-step" Procedure Combining Power and Confidence Interval Expected Width. The American Statistician 2001; 55: 288-292.
19. Cesana BM. Sample size for testing and estimating the difference between two paired and unpaired proportions: a "two-step" procedure combining power and the probability of obtaining a precise estimate. Statistics in Medicine 2004; 23: 2359-2373.
20. Lehmann EL. Testing Statistical Hypotheses, New York: Wiley, 1959.
21. Julious SA. Tutorial in Biostatistics. Sample sizes for clinical trials with Normal data. Statistics in Medicine 2004; 23: 1921-1986.
22. SAS Institute Inc. SAS^{Ò} Language: Reference, Version 6 (1 ed, chapter 11). Cary NC: SAS Institute Inc., 1990.
23. Grieve AP. Samples Sizes and Confidence Intervals. The American Statistician 44, 190.
24. Johnson NL, Kotz S, Balakrishnan N. Continuous Univariate Distributions (Vol. 1; 2nd ed). New York: John Wiley & Sons, Inc. 1994.- pp 415 - 493.
25. Hermansen MN, Nielsen KG, Buchvald F, Jespersen JJ, Bengtsson T, Bisgaard H. Acute Relief of Exercise-Induced Bronchoconstriction by Inhaled Formoterol in Children With Persistent Asthma. Chest 2006; 129: 1203-1209.
26. Kraus WE, Houmard JA, Duscha BD, Knetzger KJ, Wharton MB, McCartney JS, Bales CW, Henes S, Samsa GP, Otvos JD, Kulkarni KR, Slentz CA. Effects of the amount and intensity of exercise on plasma lipoproteins. New England Journal of Medicine 1994; 347: 1483-1492.
27. Johnson NL, Kotz S, Balakrishnan N. Continuous Univariate Distributions (Vol. 2; 2nd ed). New York: John Wiley & Sons, Inc. 1995 - pp 508 – 544.
28. SAS Institute Inc. SAS/IML^{Ò} User's Guide: Version 8. Cary NC: SAS Institute Inc., 1999.

© CIC Edizioni Internazionali