SAMPLE SIZE CALCULATION PROCEDURE
FOR PLANNED STUDIES AND ITS PROGRAM IMPLEMENTATION

Lytvynenko Y. Y., Nessonova M. N.
National University of Pharmacy, Kharkiv, Ukraine
zenyalyt@gmail.com

Introduction. Sample size determination is an important part of research planning in studies aimed at testing and confirmation of formal hypotheses. This task is particularly relevant for cost-effectiveness substantiation of the research and for its financial planning, that are integral parts of market researches, clinical trials, state public researches or studies, which are carried out under grants. The major part of methodical and teaching materials, instructions, and procedures proposed in the literature, as well as the methods implemented in specialized computer programs, focuses on planning methods of sample size evaluation for tasks of two-groups comparisons by some variable of interest, which is generally numeric. At the same time the problems of frequency assessment for qualitative feature in a single sample are insufficiently considered in the literature and implemented in specialized software.

The aim of the work is to give a survey of methods for sample size determination in a case of a dichotomous feature occurrence frequency assessment, and to introduce the software tool, developed by authors, which allows executing appropriate calculations.

Materials and methods. The computer program is written using Python 3.4.4 Integrated Development Environment (IDLE).

Results and discussion. Basic input parameters in the tasks under consideration are: confidence probability level, guess value of the expected frequency, and necessary precision of its estimate (width of confidence interval). At first the program gives user the possibility to choose one of confidence probability levels, which are most commonly used in practical researches (95%, 99% or 90).

To calculate the sample size, which is ample for expected frequency assessment, accurate within given threshold, we use formula:

$$n = \left[Z_{\alpha/2}^2 \cdot \frac{p \cdot (1-p)}{I^2} \right],$$

where n – required volume of sample; p – expected frequency of feature occurrence in the population; I – a half of confidence interval width (precision of frequency estimation); $Z_{\alpha/2}$ – standard Gaussian distribution quantile of $\alpha/2$ order, where α – type I error level (determined as 1 minus confidence probability); $[\cdots]$ – operation of rounding to the nearest greater whole number. In cases, when researcher

has no assumptions about specific value of the frequency expected, we use formula given above with $p = 0.5$. The current formula is most often cited in accessible literature sources as a single alternative to evaluate the necessary sample size for dichotomous variable expected frequency assessment in one sample. The formula is valid when the total size of the population is unknown or immense. On the other hand, in situations when we know the total size of the population, the required number of observations may be considerably less, if it is calculated by formula:

$$n = \left[\left(\frac{1}{N} + \frac{4 \cdot I^2}{Z_{\alpha/2}^2 \cdot p \cdot (1 - p)} \right)^{-1} \right],$$

where N stands for total population size.

If the researcher's task is not in obtaining the confidence interval for the frequency, but in confirmation of the hypothesis that the incidence of a feature of interest in population is not greater than some expected value, then in cases when total size of population is unknown or vast the following formula should be used to calculate the smallest necessary sample size:

$$n = \left[\frac{\ln \alpha}{\ln(1 - p)} \right].$$

For the same purpose another formula should be used if the total population size is known:

$$n = \left[\left(1 - \sqrt[Np]{\alpha} \right) \cdot \left\{ \frac{1 + N \cdot (1 - p)}{2} \right\} \right],$$

where $\{ \cdots \}$ means integer part of a number. Moreover, the last two formulae are preferable in cases when the expected frequency is close to 0 or to 1 (100%).

In the proposed software tool the particular method of sample size calculation is selected proceeding from dialogue with the user, who specifies the goals and objectives of his experiment, and inputs certain parameters he wants to assess.

Conclusions. The proposed software for sample size determination in various alternative experiments, relating to assessment of a feature incidence in the population, is supposed to be useful for clinical researchers, market research analysts, research engineers, quality control specialists and many others in their professional activity. One of prospective and challenging directions of our investigations is the software enhancement to other tasks of sampling studies planning (central tendency and variation estimation for numeric variables, among-groups comparisons of quantitative and qualitative variables, correlation power assessment, etc.). Another prospective development direction is to upgrade present software tool user interface thought the instrumentality of object-oriented programming.