PhUSE 2008

Paper ST02

Evaluation of Diagnostic Agents: a SAS Macro for Sample Size Estimation Using Exact Methods

Stefano Vezzoli, CROS NT, Verona, Italy

ABSTRACT

In a study assessing the diagnostic performance of a medical test, proper sample size planning is required to yield reliable results. For a dichotomous test, appropriate primary endpoints are represented by sensitivity and specificity. The aim of this paper is to present a SAS® macro developed to estimate the sample size needed to determine if minimally acceptable standard for the two performance measures are achieved. We show that the sample size calculation based on a normal approximation to the binomial distribution may be misleading. Exact methods are presented and implemented in the macro. This choice leads to larger sample size requirements, but allows to keep Type I error rate and power under control.

INTRODUCTION

When a diagnostic test yields a dichotomous result, four combinations of test result and disease state are possible: true positive, true negative, false positive and false negative. A true positive occurs when a diseased subject is correctly classified testing positive, and a false negative occurs when a diseased subject tests negative. Similarly, a true negative or a false positive occurs when a non-diseased subject tests negative or positive, respectively.

Table 1. Classification of test results by disease status

		True disease state	
		Present	Absent
Test result	Positive	TP (True Positive)	FP (False Positive)
	Negative	FN (False Negative)	TN (True Negative)

Sensitivity refers to how good a test is at correctly identifying subjects who have the disease. In a suitable experiment the sensitivity can be estimated as TP/(TP+FN). Specificity, on the other hand, is concerned with how good the test is at correctly identifying subjects without the condition of interest. In a suitable experiment the sensitivity can be estimated as TN/(TN+FP). Sensitivity and specificity are basic measures of performance for a diagnostic test. They each provide distinct and equally important information.

In the Points to Consider on the Evaluation of Diagnostic Agents [1] adopted by CPMP in 2001, sensitivity and specificity are designated as appropriate primary endpoints for the evaluation of the diagnostic performance of an experimental agent, when the test result is dichotomous. This recommendation is confirmed in the final version of the document Statistical Guidance on Reporting Results from Studies Evaluating Diagnostic Tests [2], released by US FDA in 2007.

We consider a non-inferiority trial aiming to show acceptable levels of inferiority when comparing the diagnostic agent with an absolute standard reference. The absolute standard by definition can truly reflect the presence or absence of the target disease [1]. The reference diagnostic procedure is frequently too expensive, hazardous, invasive or difficult to operate. A less sophisticated, safer or non-invasive test procedure, but with higher error rates, may be preferred when the loss of sensitivity and specificity is determined to be clinically acceptable. The same statistical methodology can be adopted also using a surrogate standard, such as a set of tests and clinical information that would serve as a designated reference standard if known to provide a very good approximation to the true disease state [1].

In order to design the study, some minimally acceptable sensitivity and specificity should be specified in advance by the investigators. Let ($sens_0$, $spec_0$) denote such values. Formally, the study will test the one-sided non-inferiority hypothesis

$$H_0 : \{ \text{sens} \le \text{sens}_0 \text{ or } \text{spec} \le \text{spec}_0 \} .$$

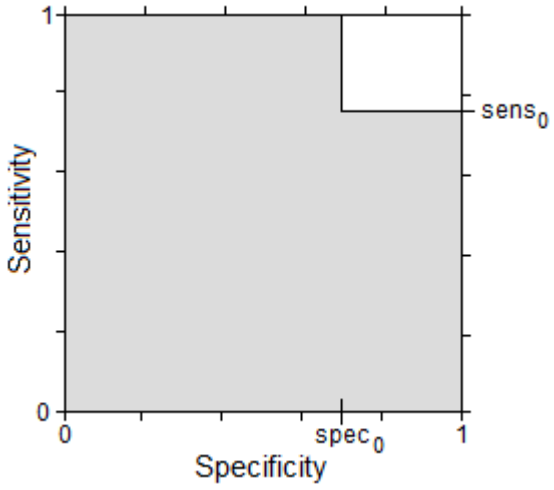
From a study that rejects H_0 it will be concluded that sens > $sens_0$ and spec > $spec_0$, i.e. that the diagnostic test meets minimal criteria. Note that the null hypothesis can also be formulated in terms of non-inferiority margins as

$$H_0 : \{ \text{sens} - 1 \le -\Delta_{sens} \text{ or } \text{spec} - 1 \le -\Delta_{spec} \},$$

assuming that both the sensitivity and specificity of the reference standard equal 1 (absence of classification error). It follows that $sens_0 = 1-\Delta_{sens}$ and $spec_0 = 1-\Delta_{spec}$.

The hypothesis can be tested by calculating a joint rectangular 1-α confidence region for (sens, spec). Following the approach of Pepe [3], we generate a rectangular confidence region made up of the cross-product of two one-sided, $1-\alpha^* = \sqrt{1-\alpha}$ confidence intervals. We are using the fact that data from diseased and non-diseased subjects, on which sensitivity and specificity are respectively estimated, are independent. If the 1-α confidence region for (sens, spec) lies entirely within the region of acceptable values (unshaded region in Fig. 1), one can reject H_0 and make a positive conclusion about the diagnostic test.

Figure 1. Regions in the (sens, spec) space for a binary test that correspond to unacceptable tests (H_0, shaded region) and acceptable tests (unshaded region). (Figure adapted from [3].)



We will show in this paper that the standard, normal-approximate approach to hypothesis testing and sample size estimation is defective in controlling Type I error rate and power, and is not to be trusted. In addition, we will motivate and illustrate an alternative exact method to deal with the problem. We will ultimately present a SAS macro for calculating the adequate sample size under this approach. The macro is available on request from the author.

SAMPLE SIZE BASED ON THE NORMAL APPROXIMATION

Sensitivity and specificity are estimated separately on diseased and non-diseased subjects. It follows that the sample sizes for the two groups (denoted with n_D and n_{ND}) must also be calculated separately. They should be chosen sufficiently large to ensure that a positive conclusion will be drawn with power 1-β if the accuracy of the test is in fact at some specified, desirable levels. We denote these desirable classification probabilities by ($sens_1$, $spec_1$). If the power of each single test is $1-\beta^* = \sqrt{1-\beta}$, then the global power 1-β is ensured [3]. Note that the sample sizes estimation procedure is exactly the same in a case-control study as in a cohort study because sensitivity and specificity condition on disease status.

From now on, we will consider the problem of sample size estimation for testing a single binomial proportion. The same considerations clearly applies equally to both the sensitivity and specificity.
In the example below, for the joint test on sensitivity and specificity, we set the significance level α at 0.1 and the power 1-β at 0.8. This means that for each single test we will have α^* = 0.05 and 1-β^* = 0.9, since

$$\alpha^* = 1-\sqrt{1-\alpha} = 1-\sqrt{1-0.1} \cong 0.05$$
$$1-\beta^* = \sqrt{1-\beta} = \sqrt{0.8} \cong 0.9 .$$

The most common method for testing a single binomial proportion is based on the asymptotic normal approximation. Relying on the asymptotic theory, the sample size can be easily calculated using the PROC POWER added to SAS/STAT® since Version 9.1. The procedure requires the user to specify the type of test, the desired significance level, whether the test is one-sided or two-sided, the parameter values under the null and alternative hypotheses, and the desired power for the test at the specified alternative. We are considering a one-sided test of a single binomial proportion with alternative greater than null value, so the ONESAMPLEFREQ statement with SIDES=U option will be used. For example, let 0.9 be the expected sensitivity or specificity, and 0.75 be the minimal acceptable level above which the 95% lower confidence limit of the expected value is required to fall with probability 0.9 (i.e. the power of the test is 90%). The following SAS code estimates the sample size needed in this situation.

```
proc power;
  onesamplefreq test=z method=normal
    sides           = u
      alpha           = 0.05
      nullproportion = 0.75
      proportion      = 0.9
      ntotal          = .
      power           = 0.9;
run;
```

Under the normal approximation, 54 subjects are needed to achieve 90% power. But if we compute the significance level of the normal-approximate test under the true binomial distribution, we find that the actual Type I error rate equals 0.0525, greater than the nominal level we required. This calculation can be performed by the PROC POWER, by specifying the option METHOD=EXACT and indicating power instead of sample size as the result parameter. One may think that increasing the sample size will result in lowering the significance level. Unfortunately, the solution is not so obvious, as it will be shown below.

Using the ODS capabilities, we can plot the achieved power (Fig. 2) and the actual Type I error rate (Fig. 3) of the test considered in the above example as a function of the sample size. What follows is the SAS code for creating the dataset (named “PlotData”) containing the variables of interest.

```
proc power plotonly;
  ods output plotcontent=PlotData;
  onesamplefreq test=z method=exact
       sides           = u
       alpha           = 0.05
       nullproportion = 0.75
       proportion      = 0.9
       ntotal          = 50
       power           = .;
  plot x=n min=50 max=100 step=1;
run;
```

Figure 2. Achieved power of the one-sided test based on the normal approximation ($\alpha^* = 0.05$, $p_0 = 0.75$, $p_1 = 0.9$).

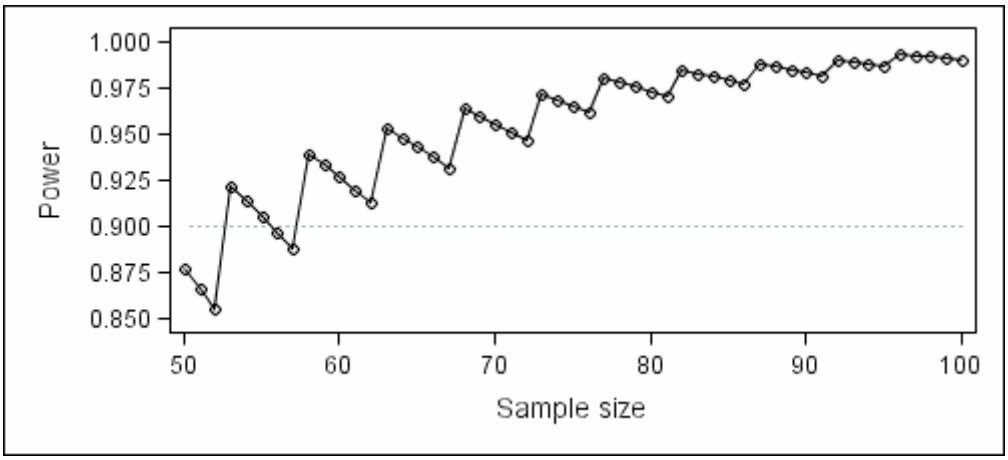
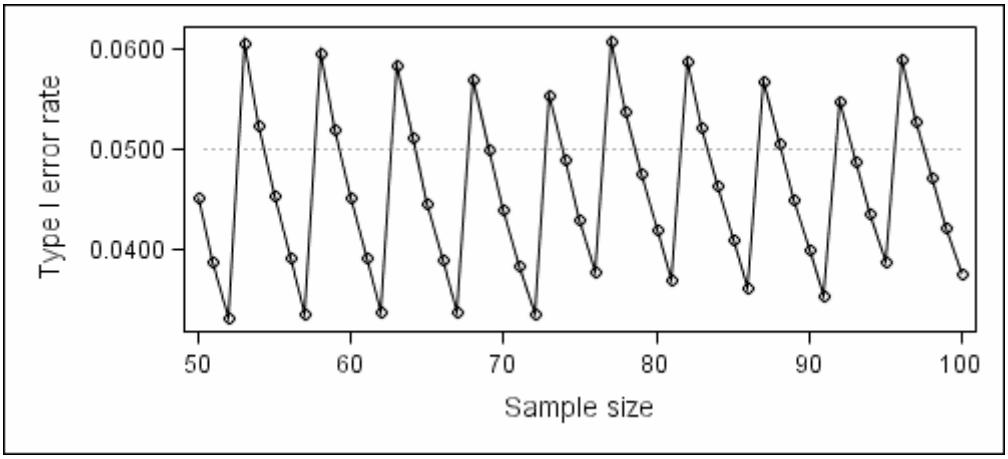


Figure 3. Actual Type I error rate of the one-sided test based on the normal approximation ($\alpha^* = 0.05$, $p_0 = 0.75$, $p_1 = 0.9$).



From the graphs we can see that:

- the power curve does not monotonically increase: it sometimes decreases with increasing sample size (Fig. 2);
- the Type I error rate fluctuates around the nominal α^* level, also for relatively large sample sizes (100 is nearly the double of the sample size previously prescribed) (Fig. 3).

These irregularities have already been largely reported [4,5]. The "saw-toothed" behavior of these functions is due to the discrete nature of the underlying binomial distribution. Hence, under the normal approximation, we must pay attention to the choice of a sample size that ensures adequate power and significance level. This problem can be partly overcome using a test based on the true binomial distribution instead of a normal-approximate test.

SAMPLE SIZE BASED ON THE BINOMIAL DISTRIBUTION

We now consider the hypothesis testing problem using the actual binomial distribution, rather than the usual normal approximation. Some authors refers to this as the "exact" procedure. However, the inference is not exact in the sense that significance level exactly equals nominal value. Rather, the nominal value is the upper bound for the true significance level. Thus in the sample size estimation we must only take care of power requirement. In the next figures, the achieved power and the actual Type I error rate under the normal approximation and the exact approach are compared. These latter values can be calculated by the PROC POWER, specifying TEST=EXACT.

Figure 4. Achieved power of the one-sided test based on the normal approximation (blue) and on the exact approach (red) ($\alpha^* = 0.05$, $p_0 = 0.75$, $p_1 = 0.9$).

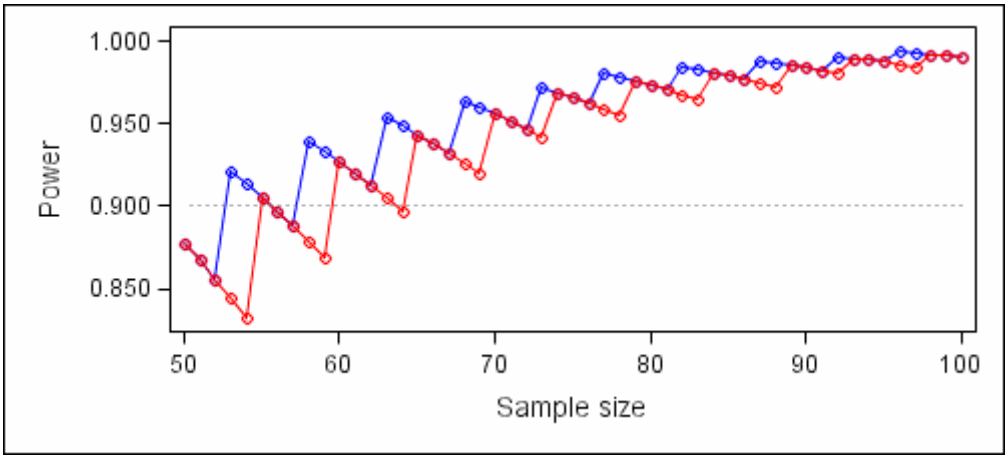
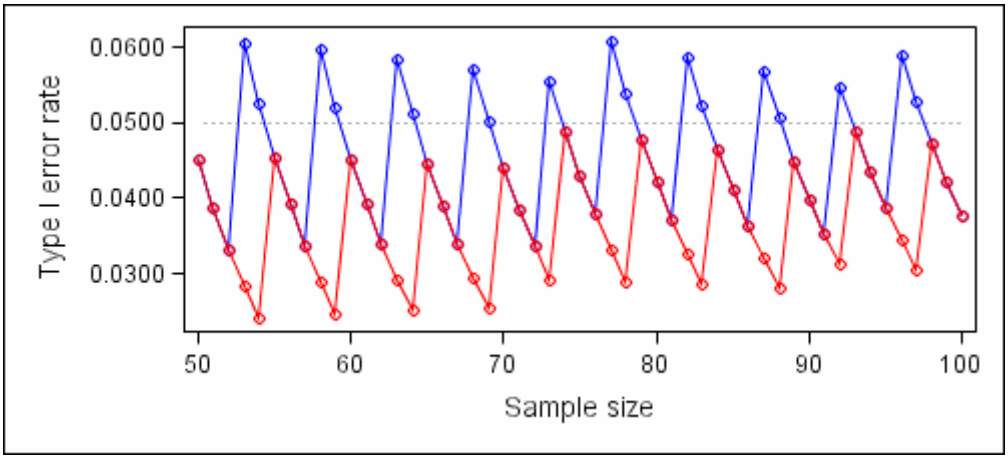


Figure 5. Actual Type I error rate of the one-sided test based on the normal approximation (blue) and on the exact approach (red) ($\alpha^* = 0.05$, $p_0 = 0.75$, $p_1 = 0.9$).



We note that whenever the significance level of the test based on the normal approximation is below the nominal level, it coincides with that of the exact test. For the same sample sizes the two tests also share a common power. These correspond to the cases of equality of critical values for the different approaches.

With PROC POWER the sample size is not available as result parameter when specifying the TEST=EXACT or METHOD=EXACT options. Instead, our macro can calculate *n* under the approach based on the exact binomial test. Under this approach the required significance level is ensured, and the shape of the power function allow us to straightforwardly determine a sample size such that the condition on the power is satisfied. Due to the discreteness and "saw-toothed" behavior of the power curve, the required sample size can be defined in two ways:

1. the minimum sample size *n* which satisfies $1-\beta_A \geq 1-\beta^*$, where $1-\beta_A$ is the achieved power;
2. the minimum sample size *n* which satisfies $1-\beta_A \geq 1-\beta^*$, and the condition is also satisfied for any sample size larger than *n*.

Various authors recommend the second, stronger condition to be satisfied [6,7]. For example, looking at the red plot in Fig. 4 we can see that $n = 55$ satisfies condition 1, while $n = 65$ satisfies condition 2.

When using the exact method, we generate the confidence intervals using the Clopper-Pearson method [8], based on the binomial distribution instead of on the normal approximation. Let *n* be the sample size, and let *x* be the number of successes (i.e. the true positives or true negatives). Therefore the $1-\alpha^*$ lower confidence bound for the true proportion is the value of *p* that satisfies

$$\alpha^* = \sum_{i=0}^{x} \binom{n}{i} \cdot p^i \cdot (1-p)^{n-i} .$$

It can be easily determined using a relation between the binomial and beta distributions, as the α^* quantile of a beta distribution Beta(x, $n-x+1$) [5]. Two-sided exact confidence intervals can be calculated using the %bnmlci macro from the Mayo Clinic [9]. Attention must be paid in the definition of the significance level: remember that we are considering a one-sided confidence region. When using the %bnmlci macro to generate the lower confidence limit, a significance level of $2\alpha^*$ must be specified.

THE %DSS MACRO

The %DSS macro is designed to estimate the sample size needed for testing the one-sided non-inferiority hypothesis on sensitivity and specificity, when comparing them to an absolute standard. The calculation is based on the exact binomial test. The macro call allows the user to set up quickly the necessary parameters needed to compute the sample size.

```
%DSS(alpha=,power=,sens=,min_sens=,spec=,min_spec=,cond=);
```

Macro %DSS contains seven parameters, six of them are required, and one (cond) is optional.

- alpha: required significance level of the test (= α);
- power: required power of the test (= $1-\beta$);

- sens: expected sensitivity (= $sens_1$);
- spec: expected specificity (= $spec_1$);
- min_sens: minimally acceptable sensitivity (= $sens_0$);
- min_spec: minimally acceptable specificity (= $spec_0$);
- cond: sample size calculated under the weak condition 1 (cond=W), the strong condition 2 (cond=S), or both the condition (cond not specified). See the previous section for details.

All the numerical parameters must be entered as proportions. A list of different values can be specified for each of these parameters (sens = 0.85 0.9 0.95, for example): the macro will calculate the adequate sample size for every possible combination of the variables. This functionality offers an easy way to assess the sensitivity of sample sizes to changes in the assumptions. The outline of the macro is as follows.

1. A check is performed on the parameter values: numerical parameters must strictly lie between 0 and 1, and minimally acceptable levels must be strictly inferior to desirable ones. Otherwise, an error message is returned.
2. The significance level α^* and the power $1-\beta^*$ for the tests on sensitivity and specificity are calculated in order to ensure a global significance level of α and a global power of $1-\beta$.
3. The sample sizes n_D and n_{ND} for the normal-approximate test are determined by using the PROC POWER, specifying the options TEST=Z METHOD=NORMAL, and indicating NTOTAL as the result parameter. The two sample sizes are calculated in the same PROC step using two different ONESAMPLEFREQ statements.
4. The power of the exact binomial test is explored for sample sizes between 0.8 and 1.5 times the approximate sample sizes calculated in the previous step. These margins are extremely conservative in order to ensure that an adequate range of sample sizes is taken into account. Using the ODS capabilities of PROC POWER with the option TEST=EXACT, datasets containing the actual significance level, the achieved power and the critical value for each sample size are generated.
5. For both the sensitivity and specificity two sample sizes are retrieved from the generated datasets, the first satisfying condition 1 and the second condition 2 as defined in the previous section. Note that it is not theoretically possible to construct an algorithm to pick the first sample size so that the power never goes below the desired level thereafter, because it would be impossible for the computer to check the infinite collection of values larger than n. However from the practical viewpoint the power function has the characteristic that it decreases slowly and then jumps up, and then cyclically repeats the decreasing trend followed by an upward jump. The jump always occurs at a higher level of power than in the previous cycle. Therefore we increase the sample size until the trough in the saw is above the required power level, and then we declare the last time the power function crossed the desired level as the prescribed sample size [4].
6. The results are presented with the PROC REPORT. Nominal and actual significance level and power, sample size under the two conditions (unless otherwise requested), and respective critical value are reported for each possible combination of macro parameters.

EXAMPLE

Suppose we wish to investigate a new diagnostic procedure with expected sensitivity of 0.9 and expected specificity of 0.95. The minimal acceptable sensitivity and specificity are 0.75 and 0.8, respectively. Thus ($sens_1$, $spec_1$) = (0.9, 0.95) and ($sens_0$, $spec_0$) = (0.75, 0.8). Conclusions will be based on a 90% confidence region using one-sided exact confidence limits at 90% power. The adequate sample size can be calculated by setting the %DSS macro parameters as follows.

```
%DSS(alpha=0.1,power=0.9,sens=0.9,min_sens=0.75,spec=0.95,min_spec=0.8,cond=);
```

After submitting the above code, we obtain the following output.

```
---------------------------------- SAMPLE SIZE CALCULATION ----------------------------------
```

	ALPHA		POWER		SENS		SPEC		DIS		NOT DIS	
Scenario	Nominal	Actual	Nominal	Actual	Min	Exp	Min	Exp	N	Crit	N	Crit
1 W	0.1	0.096	0.9	0.924	0.75	0.9	0.8	0.95	69	58	50	45
S	0.1	0.097	0.9	0.948	0.75	0.9	0.8	0.95	74	62	56	50

Since we specified only one combination of parameters, the %DSS macro evaluated only one scenario. The sample sizes are determined under the weak ("W" row) and the strong ("S" row) conditions. Focusing our attention on the calculation under the weak condition, we have n_D = 69 and n_{ND} = 50. We conclude that the diagnostic test meets the minimal criteria if the number of true positives and of true negatives are greater or equal to 58 and 45 ("Crit" columns), respectively. For this example the results are consistent with the sample sizes estimated by simulation in Pepe (n_D = 70, n_{ND} = 50) [3].

Suppose that we observed 58 true positives out of 69 diseased patients. The exact binomial confidence interval can be calculated using the %bnmlci macro from the Mayo Clinic [9], specifying a significance level of $\alpha^*/2$.

```
%let alpha=0.1;
%let width=%sysevalf(100*(2*(1-&alpha)**.5-1));

%bnmlci(width=&width,x=58,n=69);.
```

The lower confidence bound is 0.7506 > 0.75, as we expected. Applying the %bnmlci macro to the data from non-diseased patients, using the same significance level, we can generate the 90% confidence region.

LIMITATIONS AND COMPARISONS

A drawback of the proposed approach is that it may be overly conservative, since we are requiring that $\alpha_A \leq \alpha^*$ and $\beta_A \leq \beta^*$ for both the sensitivity and specificity, where the index "A" denotes the actual value. In fact, we have replaced with these stronger conditions the original ones:

$$\begin{cases} (1-\alpha_{A,sens}) \cdot (1-\alpha_{A,spec}) \geq (1-\alpha) \\ (1-\beta_{A,sens}) \cdot (1-\beta_{A,spec}) \geq (1-\beta) \end{cases}.$$

Going back to the example at the beginning of section 3, we now calculate more precisely α^* as

$$\alpha^* = 1-\sqrt{1-\alpha} = 1-\sqrt{1-0.1} = 0.0513167 .$$

Suppose that there exist two tests such that $\alpha_{A,sens} = 0.050$ and $\alpha_{A,spec} = 0.052$. The original condition on the significance level holds (0.95 x 0.948 = 0.9006 > 0.9), although the stronger condition is not satisfied by the test on specificity (0.052 > 0.051). This higher requirements in terms of Type I error rate and power may slightly increase the prescribed sample sizes.

The Stata® function *scrsize* from Fred Hutchinson Cancer Research Center [10] calculates power for the non-inferiority test on sensitivity and specificity considered in this paper. We think that the advantages of our program over the *scrsize* function are:

- the possibility of directly estimating the adequate sample sizes, avoiding the need to continue varying *n* until the desired power is reached;
- a more precise and computationally less expensive calculation of the power function, based on the relationship between the binomial and F distributions [11], rather than on simulation.

CONCLUSION

The %DSS macro allows to straightforwardly estimate the adequate sample size for testing non-inferiority of a dichotomous diagnostic test, ensuring the required Type I error rate and power. Sensitivity of sample sizes to changes in the assumptions can also be assessed. The %DSS macro is based on the PROC POWER, but add the possibility to calculate sample sizes following the exact approach.

REFERENCES

1. CPMP (2001). *Points to consider in the evaluation of diagnostic agents (CPMP/EWP/1119/98)*. EMEA, London. Available at http://www.emea.europa.eu/pdfs/human/ewp/111998en.pdf.
2. Food and Drug Administration (2007). *Statistical guidance on reporting results from studies evaluating diagnostic tests*. FDA, Rockville, MD. Available at http://www.fda.gov/cdrh/osb/guidance/1620.pdf.
3. Pepe, M. S. (2003). *The statistical evaluation of medical tests for classification an prediction*. Oxford University Press, New York.
4. Chernick, M. R. and Liu C. Y. (2002). The saw-toothed behavior of power versus sample size and software solutions: single binomial proportion using exact methods. *The American Statistician* 56: 149-155.

5. Cai, T. T. (2005). One-sided confidence intervals in discrete distributions. *Journal of Statistical Planning and Inference* 131: 63-88.
6. Borkowf, C. B. (2006). Constructing binomial confidence intervals with near nominal coverage by adding a single imaginary failure or success. *Statistics in Medicine* 25: 3679-4695.
7. Chu, H. and Cole, S. R. (2007). Letter to the editor: Sample size calculation using exact methods in diagnostic test studies. *Journal of Clinical Epidemiology* 60: 1201-1202.
8. Clopper, P. J. and Pearson E., S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* 26: 404-413.
9. Bergstralh, E. (2004). *%bnmlci: a SAS macro for calculating binomial confidence intervals*. Mayo Clinic, Division of Biostatistics. Available at http://mayoresearch.mayo.edu/mayo/research/biostat/upload/bnmlci.sas.
10. Longton, G. (2003). *scrsize: a Stata function for calculating power for a one-sample screening study with binary test outcome data*. Fred Hutchinson Cancer Research Center.
    Available at http://www.fhcrc.org/science/labs/pepe/book/prg/scrsize.ado.
11. SAS Institute Inc. (2004) *SAS/STAT 9.1 user's guide*. SAS Institute Inc., Cary, NC.

ACKNOWLEDGMENTS

I would like to thank Prof. Maria Grazia Valsecchi, Dr. Laura Antolini (University of Milan-Bicocca) and Dr. Cristina Ester Colombo (University of Milan) for their suggestions and support.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Stefano Vezzoli
CROS NT S.r.l.
Via Germania 2
Verona / 37135
Work phone: +39 045 8202666 Fax: +39 045 8205875
Email: stefano.vezzoli@cros.it Web: http://www.cros.it