A SAS® Macro for Estimating Power for ROC Curves
One-Sample and Two-Sample Cases

by Richard C. Zepp, Cleveland Clinic Foundation, Cleveland, OH

ABSTRACT

A SAS Macro for estimating the power of statistical tests involving the area under ROC curves is presented. The macro handles single reader ROC curve testing scenarios, including comparison of two areas estimated from either paired or unpaired data and comparison of a single area to a null value. The macro considers both continuous and discrete confidence scales and allows both one and two-tailed hypothesis tests.

INTRODUCTION AND PURPOSE

In medicine it is important to properly diagnose the disease status of patients. Often more than one test or technique is available as a diagnostic tool for a particular disease or abnormality, thus it is important to know which technique has the highest diagnostic accuracy.

Sensitivity and Specificity have long been the measures used to assess and compare diagnostic tests. To estimate sensitivity and specificity, a threshold at confidence is defined, whereby if a test result exceeds that threshold, the test result is defined as 'positive'; if the test result does not achieve the threshold, the result is defined as 'negative'.

Sensitivity is the proportion of positive test results among patients with disease, and specificity is the proportion of negative test results among the non-diseased patients. These measures are the favored measures because they are independent of the prevalence of disease. One drawback for these measures is that they rely on a single defined threshold. If the test responses come from continuous distributions, then the following situation may occur:

In this figure, the normal patients have a distribution of test responses and the abnormal patients have a different, but overlapping distribution of test responses. In this case, the sensitivity and specificity will be influenced by the threshold chosen. Thus if a high threshold is chosen, the sensitivity will be low, while the specificity is high. If a low threshold is chosen, then the sensitivity will be high, while the specificity is low.

ROC curves incorporate both sensitivity and specificity for the entire range of thresholds. ROC curves are constructed by graphing the sensitivity of a test versus the compliment of its specificity for all possible thresholds. ROC curves are plotted on a unit grid with the curve beginning at (0,0) (sensitivity=0 and specificity=1) and ending at (1,1) (sensitivity=1 and specificity=0). A test that is no better than random chance would pass in a straight line through (0,0) and (1,1). As the accuracy of the test increases, the ROC curve "bends" toward the point (0,1); therefore, the ROC curve of a perfectly accurate test would consist of two straight lines connecting (0,0) with (0,1) and (0,1) with (1,1).

ROC curves are used frequently in diagnostic radiology where investigators are often comparing two different modalities (e.g., CT and MRI) to determine which is better at diagnosing a certain disease or abnormality. That is, ROC curves are used to assess the diagnostic accuracy of test modalities or the ability of the modalities to separate the normal patients from the abnormal patients.

The usual method for assessing diagnostic accuracy is by examining the area under the ROC curve. This area can be interpreted as the "probability of correctly ranking a (normal, abnormal) pair" (Bamber, 1975). The area under an ROC curve can be estimated both parametrically (Metz, Wang, and Kronman, 1984) and nonparametrically (DeLong, DeLong, and Clarke-Pearson, 1988).
In an ROC study, the test results of two groups of patients (diseased and non-diseased) are compared. (Note: The unequivocal disease/non-disease status must be determined independently from the test modalities under investigation.) The test results can be summarized on either an ordinal or continuous scale. For inherently binary decisions, a score describing the confidence that the disease is present can be used. A typical ordinal confidence scale is:

1 = patient definitely normal
2 = patient probably normal
3 = patient possibly abnormal
4 = patient probably abnormal
5 = patient definitely abnormal

Continuous confidence scales (i.e., 0-100% confidence of the presence of disease) can also be applied.

The continuous scale is often preferred because 1) the parametric curve fitting algorithms are more likely to converge (Metz, 1989), 2) this scale is statistically more powerful than ordinal scales (Obuchowski, 1994), 3) it more closely resembles actual diagnostic thinking (Rockette, Gur, and Metz, 1992), and 4) studies have demonstrated no evidence of bias between the two types of scales (Rockette, Gur, and Metz, 1992 and King, Britton, Gur, Rockette, and Davis, 1993).

Asymptotic z-tests (Hanley and McNeil, 1983) are used for the comparison of ROC areas. To estimate the power of planned comparisons, an approximation of the standard error (SE) of the ROC area is needed. Approximations of the SE have been given as functions of the area under the ROC curve and the sample size (Hanley and McNeil, 1982 and Obuchowski, 1994). Thus, for a hypothesized area(s) under the curve(s) and desired sample size, the power of the associated test statistic can be determined.

The method of approximating the SE varies depending on whether an ordinal or continuous scale is used. For continuous scales, the exponential distribution based SE estimate proposed by Hanley and McNeil (Hanley and McNeil, 1982) can be used. For ordinal data, the Hanley-McNeil estimates may lead to an inflated estimate of power; thus, the binormal SE estimate proposed by Obuchowski should be substituted (Obuchowski, 1994).

When comparing the accuracy of two modalities whose ROC areas are estimated from the same patient sample, the correlation between the areas (r) must be considered. Note that the correlation between ROC areas is *not* equivalent to the correlation between test results. Hanley and McNeil (Hanley and McNeil, 1983) have developed a look-up table for r. To use the table, estimates of the correlation between the test results for the normal (rN) and the abnormal (rA) patients are needed, along with the expected area under the curve.

**MACRO ROCPOWER**

**MACRO CALL:**

```
%ROCPOWER(T1, T2, TO, NA, NN, N, PERCENT, R, ALPHA, TAILS, ORDINAL)
```

The Macro ROCPOWER calculates the power for a variety of different types of single reader ROC test-statistics, including the one-sample case (e.g., H0: T1=TO) and the two-sample case (e.g., H0: T1=T2=TO) for both paired and unpaired estimates.

**PARAMETER DEFINITIONS:**

T1, T2, and T0 are hypothesized areas under ROC curves. T0 is the area under the null hypothesis. T1 and T2 are the areas under the alternative hypothesis. For the one-sample case, only T1 and T0 are specified.

Note: Of the following four parameters, only two (either NN and NA or N and PERCENT) need to be specified.

NA is the number of abnormal patients being tested by one modality in the study. (It is assumed that the number of abnormal cases tested by each modality is the same.)

NN is the number of normal patients being tested by one modality in the study. (It is assumed that the number of normal cases tested by each modality is the same.)

N is the total sample size (both normal and abnormal patients being tested by each modality. (It is assumed that the number of cases tested by each modality is the same.)

PERCENT is the percentage of abnormal patients in the total sample size.

R is the correlation between T1 and T2 when the same patients are examined by both modalities.

ALPHA is the type I error rate. The default has been set to .05.

TAILS is an indication of whether a one tailed (e.g., H A : T1 >.75) or two tailed alternative hypothesis is desired (e.g., H A : T1 >.75). The default is set to 2.

ORDINAL is an indicator of whether a continuous (0) or ordinal (1) confidence-scale has been used. The default has been set to continuous.

**OUTPUT FROM ROCPOWER:**

The macro will echo the data from the call statement and will give the estimated power of the test.

**EXAMPLES**

**Example 1:**

Blood glucose as a predictor of Diabetes

Suppose a new test based on measured amounts of blood glucose (a continuous measure) has been developed. The makers of the test claim that it is 92% accurate in diagnosing diabetes. You want to test if the diagnostic accuracy is better than 75%. You plan to administer the test to 35 known diabetics and 50 known non-diabetics. Also you plan on using a .01 significance level.

This situation describes a one tailed hypothesis test with:

H0: Tnew <= T75
HA: Tnew > .75

What power does the test statistic have with the proposed sample size?

The macro call would be:

```
%ROCPOWER(T1=.92, TO=.75, NN=50, NA=35, ALPHA=.01, TAILS=1);
```
Hypothesis tested:
H0: T1 = .75
HA: T1 > .75

With:
T1 hypothesized at .92
Alpha = .01
Number of Abnormal Cases = 35
Number of Normal Cases = 50
Standard Error(s) estimated using the Hanley-McNeil method

The estimated Power of the test is 0.893

Example 2:
Comparing MRI and CT for detecting Stenosis in the aorta
(upaired cases)

Both MRI and CT can be used to detect stenosis in the aorta and you want to determine which imaging modality has a higher diagnostic accuracy. You expect the accuracy of both tests to be at least 70%; a difference in accuracy of 20% is clinically important. You plan to randomize a total of 200 patients to either MRI or CT (100 to each modality). The prevalence of stenosis in the population you are examining is 50. Therefore, we will assume that 50 stenotic and 50 non-stenotic patients will receive each test; thus, NN and NA are both set to 50 in the macro call.

The macro call would be:
%ROCPOWER(T1=.90, T2=.70, TO=.70, NN=50, NA=50, ORDINAL=1);

with the corresponding output:
Hypothesis tested:
H0: T1 = T2 = .70
HA: T1 < T2 or T1 > T2

With:
T1 hypothesized at .90
T2 hypothesized at .70
Alpha = .05
Number of Abnormal Cases = 50
Number of Normal Cases = 50
Standard Error(s) estimated using the Obuchowski method

The estimated Power of the test is 0.781

Example 3:
Comparing MRI and CT for detecting Stenosis in the aorta (paired cases)

Suppose that the study described in Example 2 was preformed using the same set of 100 cases for both MRI and CT and that the correlation between the two areas was estimated at .41.

The macro call would be:
%ROCPOWER(T1=.90, T2=.70, TO=.70, N=100, PERCENT=.50, ORDINAL=1);

with the corresponding output:
Hypothesis tested:
H0: T1 = T2 = .70
HA: T1 < T2 or T1 > T2

With:
T1 estimated at .90
T2 estimated at .70
Alpha = .05
Percent of abnormal patients = 50%
Total Sample Size = 100
Correlation between T1 and T2 = .41
Standard Error(s) estimated using the Obuchowski method

The estimated Power of the test is 0.935

REFERENCES


SAS is a registered trademark or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

ACKNOWLEDGEMENTS

The author wishes to thank Nancy Obuchowski, PhD for her valuable suggestions.

Comments are invited. To obtain a copy of the macro, please contact the author:

Richard Zepp
Cleveland Clinic Foundation
Department of Biostatistics and Epidemiology, P88
9500 Euclid Avenue
Cleveland, OH 44195

e-mail: rzepp@bio.ri.ccf.org