Notes on sample sizes required for Receiver Operator Characteristic (ROC) curves

14 November 2018

1 Calculating sample size needed to estimate area under an ROC curve given a specified CI and confidence interval

Following are the steps to calculating sample size needed to estimate area under an ROC curve given a specified CI and confidence interval. These steps are based on Obuchowski et al. [2004], Obuchowski [2005].

1.1 Calculating ratio of number of non-cases to cases

We first need to calculate k which is the ratio of number of non-cases to cases. This can be calculated as follows:

$$k = \frac{1 - PREV_p}{PREV_p}$$

$$where:$$

$$PREV_p = \text{prevalence of the cases in the population}$$

Given that we don't have specific values for the prevalence of children demonstrating appropriate infant and young child feeding, we make an educated guess of what this prevalence can be. It is likely that the practice of appropriate infant and young child feeding is low and is probably at best at the 20% level. Demographic and Health Survey (DHS) for Burkina Faso in 2010 show that for the whole country in total, minimum acceptable diet (MAD) was at 3.1% and about 2.4% for rural areas. In 2016, SMART surveys conducted in parts of Burkina Faso have shown in increase in this indicator to about 20%.

Using this, we estimate k as follows:

$$
\begin{aligned}
k &= \frac{1 - PREV_p}{PREV_p} \\
&= \frac{1 - 0.2}{0.2} \\
&= 4
\end{aligned}
$$

1.2 Calculating the binormal distribution parameter

We then need to calculate the binormal distribution parameter A. A is one of two (the other being B) parameters that describe the ROC curve. However, these parameters are rarely empirically estimated as they represent unobserved binormal variables. However, for sample size calculations, parameter A can be estimated based on an expected area under the curve (AUC) of the ROC curve. The AUC corresponds to the expected level of agreement or accuracy of the metric being tested with the actual status or condition observed.

For our purposes, the AUC will be for how much agreement or accuracy there is for MAD and/or good ICFI to predict mean nutrient and mean energy adequacy. # Given AUC, A can be calculated as follows:

$$
\begin{aligned}
A &= \phi^{-1}(AUC) \times 1.414 \\
\\
where &: \\
\\
\phi^{-1} &= \text{inverse of the cumulative normal disribution function} \\
\\
AUC &= \text{expected area under the curve}
\end{aligned}
$$

Since we do not have prior knowledge of the AUC from previous studies, we set AUC at 0.90 which is the AUC value we would assume that would show best agreement between ICFI and indicators for nutrition and energy adequacy. This gives us the following A parameter:

$$
\begin{aligned}
A &= \phi^{-1}(AUC) \; \times \; 1.414 \\
&= \phi^{-1}(0.90) \; \times \; 1.414 \\
&= 1.281552 \; \times \; 1.414 \\
&= 1.812114
\end{aligned}
$$

1.3 Calculating the variance function

To be able to continue with the calculations, we will need to calculate the variance function (VF) of the accuracy estimate of the ROC curve analysis. This variance function (VF) can be expressed using the A parameter and k which we have previously calculated as shown below:

$$VF \; = \; 0.0099 \; \times \; e^{-A \; \times \; A/2} \; \times \; \left[(5 \; \times \; A^2 \; + \; 8) \; + \; \frac{(A^2 \; + \; 8)}{k} \right]$$

$$
\begin{aligned}
&where : \\
&A \; = \; 1.812114 \\
&k \; = \; 4
\end{aligned}
$$

Using the values calculated for A and k previously, we calculate VF as follows:

$$
\begin{aligned}
VF \; &= \; 0.0099 \; \times \; e^{-A \; \times \; A/2} \; \times \; \left[(5 \; \times \; A^2 \; + \; 8) \; + \; \frac{(A^2 \; + \; 8)}{k} \right] \\
&= \; 0.0099 \; \times \; 0.193616 \; \times \; \left[(5 \; \times \; 3.283757 \; + \; 8) \; + \; \frac{(3.283757 \; + \; 8)}{4} \right] \\
&= \; 0.001916798 \; \times \; [\; 24.41878 \; + \; 2.820939 \;] \\
&= \; 0.001916798 \; \times 27.23972 \\
&= \; 0.05221305
\end{aligned}
$$

1.4 Calculating number of cases needed in the study sample

Now we can calculate the number of cases (i.e. those who have the condition of interest). For the case of ICFI or IYCF, this will be those children who exhibit or demonstrate appropriate infant and young child feeding. This can be calculated as follows:

$$N \ = \ \frac{Z^2_{\alpha/2} \ \times \ VF}{L^2}$$

$$where:$$

$$Z_{\frac{\alpha}{2}} \ = \ 1.96 \text{ for a 95\% CI}$$

$$L \ = 0.05 \text{ (desired half-width of the CI)}$$

$$VF \ = \ 0.05221305$$

Using the value for VF calculated previously, we calculate N as follows:

$$\begin{aligned} N \ &= \ \frac{Z^2_{\alpha/2} \ \times \ VF}{L^2} \\ &= \ \frac{1.96^2 \ \times \ 0.05221305}{0.05^2} \\ &= \ 80.23268 \ \approx \ 81 \end{aligned}$$

1.5 Total sample size

The total sample size n needed for the ROC analysis can then be calculated as follows:

$$n \ = \ N \ \times \ (1 \ + \ k)$$

$$where:$$

$$N \ = \ 81$$

$$k \ = \ 4$$

Using the N and k values calculated earlier, n is calculated as follows:

$$
\begin{aligned}
n &= N \times (1 + k) \\
&= 81 \times 5 \\
&= 405
\end{aligned}
$$

We will need a sample size of about **405** children about **81** of which will be children who are practising appropriate infant and young child feeding.

References

Nancy A Obuchowski. ROC Analysis. *American Journal of Roentgenology*, 184(2):364–372, feb 2005. ISSN 0361-803X. doi: 10.2214/ajr.184.2.01840364. URL https://doi.org/10.2214/ajr.184.2.01840364.

Nancy A Obuchowski, Michael L Lieber, and Frank H Wians. ROC Curves in Clinical Chemistry: Uses, Misuses, and Possible Solutions. *Clinical Chemistry*, 50(7):1118–1125, 2004. ISSN 0009-9147. doi: 10.1373/clinchem.2004.031823. URL http://clinchem.aaccjnls.org/content/50/7/1118.