A graphic-based approach to obtain sample size in reliability studies

F. Ingravalle, G. Ru, M. Caramelli

CEA - Istituto Zooprofilattico Sperimentale del Piemonte, Liguria e Valle d'Aosta

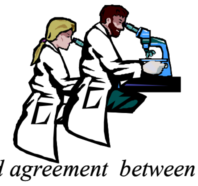
Introduction

Reliability studies are usually designed to evaluate the agreement between results obtained both by the application of the same method by the same rater in different times (repeatability) and by the use of the same diagnostic assay in similar conditions, by different raters or laboratories (reproducibility).

Cohen's k is the most suitable measure of reliability when the results are categorical (see fig. 1).

Fig. 1: Cohen's k statistic (see Fleiss [1]):

$$k = \frac{(p_o - p_e)}{(1-p_e)}$$



p_o = *proportion of observed agreement between raters*

p_e = *proportion of expected agreement (due to random answers) between raters*

Aim

Although the topic of reliability has gained much attention in the literature, investigations into sample size requirements remain scarce. Aim of this study is to **propose a graphical procedure based on Monte Carlo simulations** in order to determine the necessary sample size in reliability studies for binary diagnostic assessments.

Material and methods

During the planning of a binary reproducibility study, researchers have to define the number of samples $\boldsymbol{n}$ that the 2 raters will test with the same diagnostic method. To calculate n the investigator must provide:

- a guess of the expected value of agreement ($\boldsymbol{k_{exp}}$), based on previous experience;
- a guess of the proportions of positive samples detected by each rater: $\boldsymbol{p_1}$ and $\boldsymbol{p_2}$ (in this study p_1 and p_2 are equals: $p_1 = p_2 = \boldsymbol{p}$);
- the desired confidence level in term of $\boldsymbol{\Delta k}$, *i.e.* the gap between k_{exp} and its $(1-\alpha)\%$ lower limit of the confidence interval.

In this work the formula:

$$n = \frac{(z_{\alpha/2})^2}{\Delta k^2}(1-k_{exp})[(1-k_{exp})(1-2k_{exp}) + \frac{k_{exp}(2-k_{exp})}{2p(1-p)}$$

(Shoukri [2]) is used.

The number of all the combinations of the factors ($k_{exp,}$ p, Δk) above described is infinite. In order to obtain a representative sample of these combinations, in the current study the factors were considered as random variables:

- $\boldsymbol{K_{exp}}$ is a uniform random variables from 0.5 to 0.99;
- $\boldsymbol{p}$ is a discrete random variable with minimum value equal to 5%. The other values are multiple of 5%, up to a maximum, equal to 95%;
- $\boldsymbol{\Delta k}$ is a uniform random variable from 0.01 to k_{exp}.

In this manner, sample size n, calculated applying the previous formula, and lower limit $\boldsymbol{k_{low}}$ of the $(1-\alpha)\%$ confidence interval for k_{exp} have been obtained with random assumptions.

A Monte Carlo simulation program, developed in STATA 8.2, repeated the random procedure **15.000 times**. Then weighted means of k_{low} for each value of p, n, k_{exp} wereù calculated. An example of the simulations performed is shown in fig. 2.

Finally the k_{exp}s were grouped in 10 intervals of values (from $k_{exp} \geq 0.50$ to $k_{exp} < 0.55$, from $k_{exp} \geq 0.60$ to $K_{exp} < 0.65$,…, $k_{exp} \geq 0.95$ to $k_{exp} < 1$) and a graph for each p has been built

Fig. 2: STATA8.2 browse on simulations

	p	n	kexp_class	lowexp_mean	kexp	widthexp	lowexp
1	.05	40	kexp>=0.50 & kexp<0.60	0.00	.59	.59	0
2	.05	41	kexp>=0.50 & kexp<0.60	0.01	.59	.58	.01
3	.05	43	kexp>=0.50 & kexp<0.60	0.01	.58	.57	.01
4	.05	44	kexp>=0.50 & kexp<0.60	0.01	.59	.56	.03
5	.05	45	kexp>=0.50 & kexp<0.60	0.02	.58	.56	.02
6	.05	46	kexp>=0.50 & kexp<0.60	0.03	.59	.55	.04
7	.05	47	kexp>=0.50 & kexp<0.60	0.03	.59	.54	.05
8	.05	48	kexp>=0.50 & kexp<0.60	0.04	.58	.54	.04
9	.05	49	kexp>=0.50 & kexp<0.60	0.03	.57	.54	.03
10	.05	50	kexp>=0.50 & kexp<0.60	0.03	.58	.53	.05
11	.05	51	kexp>=0.50 & kexp<0.60	0.04	.57	.53	.04
12	.05	52	kexp>=0.50 & kexp<0.60	0.02	.54	.53	.01
13	.05	53	kexp>=0.50 & kexp<0.60	0.05	.53	.53	0
14	.05	54	kexp>=0.50 & kexp<0.60	0.03	.54	.52	.02
15	.05	55	kexp>=0.50 & kexp<0.60	0.06	.53	.52	.01

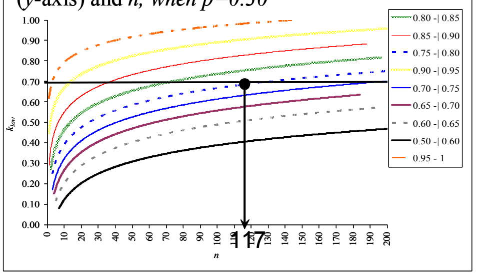
Results and conclusions

Since p can assume 19 different values (0.05, 0.10,…,0.95), 19 different graphs were obtained. In the graphs, a curve was reported for each interval of values of k_{exp} whereas the x-axis is the sample size n and *the* y-axis is the lower limit $\boldsymbol{k_{low}}$.

Figure 3 shows the graph obtained for p=0.3. After choosing values for k_{exp} and its $\boldsymbol{k_{low}}$ the researcher can find the sample size directly from the graph.

An example: a researcher is planning a reliability study for a diagnostic tool.
A good reliability is expected (k_{exp}=0.80). The researcher decides to accept $\boldsymbol{\Delta k}$ **=** 0.1 and therefore k_{low}= (0.80-0.10) = 0.70. Moreover, he assumes that the raters involved in the study, will identify as positive about p=30% of the samples.
In order to obtain n, with a 90% confidence level (ie $z_{\alpha/2}$=1.64).
the researcher draws an horizontal line at k_{low}=0.70 on y-axis and casts the point of intersection with the k_{exp} =0.80 curve on the x-axis: he will get a sample size of 117.

Fig. 3: relation between k_{exp} (curves), k_{low} (y-axis) and n, *when p=0.30*



As shown, the graphs are easy to manage and ready to use during study design.
Moreover the researcher can evaluate how n vary as k_{exp} and its lower limit change.

Bibliography

[1] Fleiss J.L., Levin B., Cho Paik M., "Statistical methods for rates and proportions", 3rd ed. John Wiley & Sons, New York, 2003

[2] Shoukri M.M., "Measures of interobserver agreement", ed. Chapman & Hall/CRC, New York, 2004