Sample size for estimating the ratio of two means

Carlotta Galeone[1,2], Angiola Pollastri[3]
[1] Istituto di Statistica Medica e Biometria "G.A. Maccacaro",Università di Milano
[2] Dipartimento di Epidemiologia, Istituto "Mario Negri", Milano
[3] Dipartimento di Metodi Quantitativi per le Scienze Economico-Aziendali, Università di Milano-Bicocca.

e-mail: angiola.pollastri@unimib.it

Keywords: ratio of two means, sample size determination, distribution of two correlated Normals.

1. Introduction

The decision about the sample size is particularly difficult when we have to estimate the ratio of two means. This abstract presents a procedure for the sample size determination in the considered situation.

Let us consider a bivariate random variable $(r.v.)$ (X_1, X_2) having

$$E(X_1)=\mu_1 \;;\; E(X_2)=\mu_2 \;;\; Var(X_1)=\sigma^2{}_1 \;;\; Var(X_2)=\sigma^2{}_2 \;;\; Corr(X_1,X_2)=\rho\,.$$

Let us suppose we have drawn a sample of *n'* elements and we have obtained the observations (x_{1i}, x_{2i}) *(i=1,…,n').*

If (X_1, X_2) is a Bivariate Correlated Normal (B.C.N.) or if *n'* is large, the r.v. $(\overline{X}_1, \overline{X}_2)$ tends to a B.C.N. having parameters

$$E(\overline{X}_1)=\mu_1 \;;\; E(\overline{X}_2)=\mu_2 \;;\; Var(\overline{X}_1)=\frac{\sigma^2{}_1}{n'} \;;\; Var(\overline{X}_2)=\frac{\sigma^2{}_2}{n'} \;;\; Corr(\overline{X}_1,\overline{X}_2)=\rho\,.$$

We can compute the maximum likelihood (M.L.) estimates of the means μ_1 and μ_2 indicated respectively by $\overline{x}_1$ and $\overline{x}_2$. The M.L. estimates of σ_1^2, σ_2^2 and ρ are respectively given by

$$s_1^2=\sum_{i=1}^{n'}(x_{1i}-\overline{x}_1)^2/n' \qquad s_2^2=\sum_{i=1}^{n'}(x_{2i}-\overline{x}_1)^2/n'$$

$$r=\sum_{i=1}^{n'}(x_{1i}-\overline{x}_1)(x_{2i}-\overline{x}_2)\Big/\sqrt{\sum_{i=1}^{n'}(x_{1i}-\overline{x}_1)^2\sum_{i=1}^{n'}(x_{2i}-\overline{x}_2)^2}$$

Keeping into account the result obtained by Aroian (1986) and Oksoy et Aroyan (1986) about the Distribution Function (DF) of the ratio of two correlated Normals, we can obtain the distribution of the r.v. $W_n = \frac{\bar{X}_1}{\bar{X}_2}$, indicated by $F_{W_n}(w)$. By substituting in it the estimates of $\mu_1, \mu_2, \sigma_1, \sigma_2$ and ρ, we can estimate the DF of W_n

$$\hat{F}_{W_n}(w) = L\left(\frac{a_n - b_n t_w}{\sqrt{1+t_w^2}}, -b_n, \frac{t_w}{\sqrt{1+t_w^2}}\right) + L\left(\frac{b_n t_w - a_n}{\sqrt{1+t_w^2}}, b_n, \frac{t_w}{\sqrt{1+t_w^2}}\right)$$

where

$$a_n = \sqrt{\frac{n}{1+r^2}}\left(\frac{\bar{x}_1}{s_1} - r\frac{\bar{x}_2}{s_2}\right) \qquad b_n = \sqrt{n}\left(\frac{\bar{x}_2}{s_2}\right) \qquad t_w = \sqrt{\frac{1}{1+r^2}}\left(\frac{s_2}{s_1}w - r\right)$$

and where, according to the indication of Kotz et al (2000),

$$L(h,k,\rho) = \frac{1}{2\pi\sqrt{1-\rho^2}} \int_h^\infty \int_k^\infty \exp\left\{-\frac{1}{2\left(1-\rho^2\right)}(x_1^2 - 2\rho x_1 x_2 + x_2^2)\right\} dx_1 dx_2$$

The estimate of the DF can be computed as a function of *w* using Fortran+IMSL library or MATLAB or other libraries, especially those containing routines regarding the DF of a B.C.N. or other functions which can give the DF of the B.C.N..

2. Procedure for the determination of the sample size

If the researcher wishes to control the difference between the estimate and the real value $R = \frac{\mu_1}{\mu_2}$, the criterion may be

$$P\{|W_n - R| \le \varepsilon\} = 1 - \alpha$$

where $\varepsilon > 0$ and $(1-\alpha)$ are fixed by the researcher, that is

$$P\{W_n - \varepsilon \le R \le W_n + \varepsilon\} = 1 - \alpha \tag{1}$$

The confidence intervals for R (Galeone, 2007) are given by

$$P\{W_{\alpha/2} \le R \le W_{1-\alpha/2}\} = 1-\alpha \qquad (2)$$

where $W_{\alpha/2}$ and $W_{1-\alpha/2}$ are the estimators (see Galeone et Pollastri, 2008) of $(\alpha/2)th$ and the $(1-\alpha/2)th$ quantile of the distribution of the r.v. W_n.

Comparing (1) with (2), we obtain

$$W_n - \varepsilon = W_{\alpha/2}$$

After the pilot sample or past experiencies, we have an estimate of *R*, that is $\frac{\overline{x}_1}{\overline{x}_2}$, and we must find the value of *n* such that

$$\frac{\overline{x}_1}{\overline{x}_2} - \varepsilon = c \cong w_{\alpha/2}$$

and

$$\frac{\overline{x}_1}{\overline{x}_2} + \varepsilon = d \cong 1 - w_{\alpha/2}$$

This is equivalent to finding the value of *n* such that

$$P\{W_n \le c\} = \alpha/2 \qquad (3)$$

Giving to *n* the initial value of *n'* then increasing by a unit, until the two relations are verified:

$$P\{W_n \le c\} \le \alpha/2 \qquad (4)$$

$$P\{W_n \le d\} \ge 1-\alpha/2 \qquad (5)$$

The corresponding value of *n* is the sample size required in order to guarantee the precision (1). Note that the value of *n* which satisfies the relation (4) and (5) is unique and therefore it is sufficient to find the value of *n* for one of the two relations. In Table 1, we show some examples of the determination of *n* for some values of the parameters. It is interesting to note that the value of the sample size depends on the variability of X_1 and X_2 and on the value of ε but also that the value of *n* decreases considerably when, coeteris paribus, the value of the correlation coefficient increases.

Table 1: values of sample size in different situations having fixed $(1-\alpha)=0.95$

μ_1	μ_2	σ_1^2	σ_2^2	ρ	ε	n
1	1	0.3	0.2	0.9	0.01	2190
1	1	0.3	0.2	0.5	0.01	8644
1	1	0.3	0.2	0	0.01	11525
1	1	0.3	0.2	0.9	0.1	22
1	1	0.3	0.2	0.5	0.1	87
1	1	0.3	0.2	0	0.1	116
1	2	0.1	0.2	0.9	0.05	11
1	2	0.1	0.2	0.5	0.05	29
1	2	0.1	0.2	0	0.05	39

3. Conclusions

In the present abstract, considering the information based on a pilot sample or past experiencies and the distribution of the ratio of two correlated Normals, we propose a procedure in order to estimate the sample size when the estimation of two means is required. From some examples it is possible to show that the sample size decreases if the correlation coefficient increases.

References

Aroian L.A. (1986) The distribution of the quotient of two correlated random variables. *Proceedings of the Am. Stat. Ass. Business and Economic Section.*

Galeone C. (2007) On the ratio of two Normal r.v., jointly distributed as a Bivariate Normal, *PhD Thesis, Università di Milano-Bicocca*.

Galeone C., Pollastri A. (2008) Estimation of the quantiles of the ratio of two correlated Normals, *XLIV Riunione scientifica della Società Italiana di Statistica.*

Kotz S., Balakrishnan N., Johnson N.L. (2000) *Continuous Multivariate Distributions,* Wiley, New York.

Oksoy D., Aroian L.A. (1986) Computational Techniques and examples of the density and the distribution of the quotient of two correlated normal variables, *Proceeding of the American Statistical Association Business an Economic Section.*