Management & Marketing (2010) Vol. 5, No. 1, pp. 127-134.

POSSIBLE ERRORS IN ESTABLISHING THE SAMPLE DIMENSIONS OF SURVEY INVESTIGATIONS

Alexandru ISAIC-MANIU, Irina M. DRĂGAN
Academy of Economic Studies, Bucharest, Romania

Abstract. *Starting from the „rough" use of 1,500-1,800 subject samples in field researches with populations of the dimensions of a medium town, a capital or an entire country, which generates doubts and possible questions related to the representativeness of the investigated samples as well as to the conclusions of these researches, we developed the reconstruction of the necessary interactions in determining the sample volume, and also identified some inadvertencies. Some improvements solutions were also recommended.*

Keywords: confidence interval, distribution laws, probable error, representativeness, sample volume.

1. Sample volume estimation

In organizing a survey research, one of the main problems to solve is its rational dimension. It is true that the size of volume sample *n* – by virtue of the law of large numbers – increases the precision of the results and reduces the probable average error. Taking into consideration economical criteria, it is necessary that this volume be as small as possible. Considering both aspects, we determine the minimum number of observation units that satisfies precision and safety requirements formulated in relation to the research.

In the theory and practice of surveys, „large" and „reduced volume" samples are used, depending on the homogeneity degree and on the volume of the origin general population. The calculus and interpretation of the representativeness error are determined differently: for large volume samples, the Laplace normal distribution, while for reduced volume samples, Student distribution.

The sample volume calculus is achieved starting from the permitted maximum limit error, which in case of the repeated survey, is:

$$e_{\bar{x}} = z_{\alpha} \frac{\sigma}{\sqrt{n}} \tag{1}$$

from where we can highlight *n*:

$$e_{\bar{x}}^2 = z_{\alpha}^2 \frac{\sigma^2}{n} \tag{2}$$

and thus:

$$n = \frac{z_{\alpha}^2 \cdot \sigma^2}{e_{\bar{x}}^2} \tag{3}$$

In the case of the repeated survey, the relation becomes:

$$n = \frac{z_\alpha^2 \cdot \sigma^2}{e_{\bar{x}}^2 + \frac{z_\alpha^2 \cdot \sigma^2}{N}} \quad (4)$$

In practice, the values resulted from this relation are majored in order to obtain an integer number, as well as to not fail the research, taking into account the fact that a number of questionnaires are usually rejected.

So, in order to rationally dimension the sample volume *n*, the following elements are necessary:

- ***admissible limit error*** $e_{\bar{x}}$ which is established depending on the concrete features of the problem to be solved, and on the required precision;
- ***confidence probability*** $1-\alpha$ **,** practically close enough to certainty;
- ***variance***, the σ^2 characteristic in the general population, or its estimator, established on the base of a survey;
- in the case of unrepeated sampling, the general population ***volume*** *N*, is also necessary.

In these conditions, for the most usual probability values, probability coefficients equivalences, and admissible maximum errors, we have the values of the sample volume, in Table 1.

Table 1

Values for volume *n* sample, for $e_{\bar{x}}$(%)=1-5% and P= 90-99,9%

P(%) / $e_{\bar{x}}$(%)	P=99,9% $Z_{0,001}$=3,29	P=99% $Z_{0,01}$=2,57	P=97,5% $Z_{0,025}$=2,24	P=95% $Z_{0,05}$=1,96	P=90% $Z_{0,10}$=1,645
1	27.050	16.520	12.550	9.600	6.800
2,5	4.328	2.643	2.008	1.536	1.088
3	3.006	1.836	1.394	1.067	756
5	1.082	661	502	384	272

The data confirm the fact that in the most usual conditions imagined by research and survey companies, the numbers regarding the investigated survey volume are placed around 1,500-1,800 units. A few observations regarding the calculus and results are required.

1.1. Observation I

- In relations (3) and (4), variance is considered unknown and is estimated through the maximum variance of the binary characteristic:

$$\sigma^2 = p \cdot q = p(1-p)$$

- As known, the mean of the binary characteristic belongs to the [0; 1] interval. For the most probable value of the mean equal to 0.5, we have a value of the variance equal to 0.25.
- If this method is justifiable in the calculus of sample volume in the absence if previous information regarding the population under research, the same value (0.5) should be used for the mean in determining the probable error; the values of 1%-5% of this error indicate admissible deviations of the estimated parameters from the true but unknown values of the general population. In such a situation, the sample volume results in a range of 5,000-10,000 units, different from the ones usually used by survey or marketing institutes.

Details on input indicators and results achieved are found in Table 2.

Table 2

Values for sample volume *n* and a mean *m* = *0.5*

$e_{\bar{x}}$ / P	P=99,9% $Z_{0,001}$=3,29	P=99% $Z_{0,01}$=2,57	P=97,5% $Z_{0,025}$=2,24	P=95% $Z_{0,05}$=1,96	P=90% $Z_{0,1}$=1,645
1% — 0,01 — 0,01x0,5=0,005 — 0,005² =0,000025	108.200	66.080	50.200	38.400	27.200
2,5% — 0,025 — 0,025x0,5=0,0125 — 0,0125² =0,0001525	17.318	10.576	8.035	6.146	4.354
3% — 0,03 — 0,03x0,5=0,015 — 0,015² =0,000225	12.022	7.342	5.578	4.267	3.022
5% — 0,05 — 0,05x0,5=0,025 — 0,025² =0,000625	4.328	2.643	2.008	1.536	1.088

The values of the normal standardized variable for the one-tailed case are used, although the confidence interval for the estimated parameters is two-tailed. The values of probabilistic coefficients are presented in Table 3, in the two testing alternatives.

Table 3

The values for Z_α variable

α	$\alpha = 1 - 2 \cdot \phi(Z)$	$\alpha = 0.5 - \phi(Z)$
0,001	$Z_{0,001}$=3,29	$Z_{0,001}$=3,09
0,01	$Z_{0,01}$=2,57	$Z_{0,01}$=2,33
0,025	$Z_{0,025}$=2,24	$Z_{0,025}$=1,96
0,05	$Z_{0,05}$=1,96	$Z_{0,05}$=1,645
0,10	$Z_{0,10}$=1,645	$Z_{0,10}$=1,28

This mistaken option also leads to distortions in the values of survey indicators.

1.3. Observation III

Another fundamental hypothesis, supposedly confirmed, but seldom tested, is the normality of the data (Gauss-Laplace law). The abusive use of the normal law is due to the fact that most of the times there is no preoccupation for preliminary testing in order to validate or to not validate the distribution law, and therefore final results may be corrupted in the un-normality case of data obtained through survey.

The usual Gauss-Laplace model, from a mathematical point of view is a continuous statistic distribution defined by the expression:

$$F(x;\mu;\sigma^2)=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{x}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]dx$$

where:

$$\mu\in R,\ \sigma>0,\ x\in R.$$

Therefore, the confidence interval for the mean results:

$$P\left(\left|\bar{x}-m\right|<e_{\bar{x}}\right)=1-\alpha$$

$$P\left(\bar{x}-Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\le m\le\bar{x}+Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)=1-\alpha$$

If the sample volume was determined based on the hypothesis of binary characteristic and variance (along with the previous observations), the confidence interval is usually established by introducing a new methodological discordance, for supposed normally distributed measurable characteristics.

1.4. Observation IV

In many practical situations, the *t* probabilistic coefficient is used, which has values different from the ones of Laplace variable, besides the fact that its values are dependent on the number of degrees of freedom, related to the volume of the origin population, respectively of the sample.

The *t* distribution is also known as Student distribution after the name of the mathematician statistician William Sealy Gosset (1876-1937), who was mostly known under the pseudonym *Student* and an important representative of the Anglo-Saxon statistic school. He led the laboratory of beer factory (Park Royal) using statistic procedures in interpreting the results of the analyses. He also worked at the London University College and collaborated with R.A. Fisher and K. Pearson. The original results in the theory of probable error of the mean, in reduced volume surveys, are found in the „t law". The *t* distribution is the probability distribution of a continuous random variable, with a probability density function given by:

$$f(t;\nu)=\frac{1}{\sqrt{\pi\nu}}\left(\frac{\Gamma\left[\frac{\nu+1}{2}\right]}{\Gamma\left(\frac{\nu}{2}\right)}\right)\left(\frac{1}{\left[1+\frac{t^2}{\nu}\right]^{\frac{\nu+1}{2}}}\right),$$

where $-\infty < t < +\infty$ with the parameter $\nu = 1,2,...$ and Γ is the *Gamma* function.

If the X random variable has a Student distribution with $\nu > 2$ degrees of freedom, then:

$$M(X)=0;\ Var(X)=\frac{\nu}{\nu-2}$$

The use of the Student variable values in calculus, without appropriate justification, is another source of errors that might influence the final results.

2. A few solutions

Performing survey investigations in conditions of statistic fidelity able to assure high representativeness and accuracy to the final results must correspond to some rules such as:

2.1. The dimensioning of sample volume, in the case of binary characteristics, is achieved by using the binomial or hypergeometric model, depending on the extraction type (with or without replacement), respectively the dimension of the reference population (reduced volume or high volume).

a) Binomial distribution (formulated by J. Bernoulli)

When the values of a random variable X are the proportion of A or *non A* elements (for example the proportion of citizens who intend to vote respectively who do not intend to), the typical values are:

- mean: $E(X)=np$;
- variance: $Var(X)=npq$;
- mode: $np-q \le Mo \le np+p$.

where: p and q are the elements of A and *non A*, and n – the sample volume.

The theoretical model considers there is the same probability for the occurrence of the $A_1, A_2, ..., A_n$ events, and therefore the probability for the occurrence of x out of n events is:

$$P(X=x)=C_n^x p^x q^{n-x}=\frac{n!}{x!(n-x)!}p^x q^{n-x},$$

where x is the frequency of the event under observation, with values $x = 0,1, ... ,n$.

b) The hypergeometric distribution

It is used in investigations as model especially for the description of reduced volume samples operations, with replacement sampling.

The probability that *k* favorable results be obtained out of *n* units, is determined as:

$$P\left(N;n,x\right)=\frac{C_k^x C_{N-k}^{n-k}}{C_N^n};\quad x=0,1,2,\ldots,\left(n,x\right)$$

The typical values of this distribution are:

- mean $M\left[X\right]=np$;
- variance $D^2\left[X\right]=np\left(1-p\right)\frac{N-n}{N-1}$.

when *N* is large enough, practically when $n/N<1/10$, the hypergeometric law can be approximated through the binomial law.

2.2. Establishing the main characteristic for which the research is designed; the calculations are performed for this characteristic which assures the representativeness of the entire investigation, the rest of the characteristics becoming subordinated.

2.3. In the case of binary characteristics, that is in the case of alternative questions, designing the survey parameters must be accomplished by using the binomial law model, if the survey is unrepeated, or if the population under investigation has a high volume, while in the case of comeback sampling or reduced volume populations, the hypergeometric law model is recommended.

2.4. If designing the sample volume was based on binary characteristic, then the confidence interval as well as the generalization of the results, must be accomplished based on the same hypotheses; if the essential characteristic is measurable and continuous, as are most in sociologic questionnaires, then the sample design as well as expanding the results will be performed in the hypothesis that must be tested in regard of validating the normality.

2.5. Estimating variance

a. For binary characteristics, the use of the maximum variance (*pq* = 0.25) can be chosen only as a temporary solution, and only in the case of a first research over a population. Later on, based on supplementary information, the survey estimator can be used, which gets closer and closer to the unknown variance of the general population with each research.

b. A specific survey can also be organized, with no restrictions regarding representativeness, on a reduced number of subjects, and on the basis of the values registered for characteristic *X*:

$$X=x_1,x_2,\ldots,x_n$$

The corrected survey variance can be estimated:

$$\hat{S} = \frac{\sum\left(x_i - \bar{x}\right)}{n-1}$$

which can be used as reference element in the calculus of the sample volume for the first research, whose quality shall be improved after further information.

c. If establishing the sample is compulsory, the variance can be estimated through the maximum variance of the measurable characteristic:

$$\sigma^2_{\max} = \frac{\left(x_{\min} - \bar{x}\right)^2 + \left(x_{\max} - \bar{x}\right)^2}{2}$$

d. This issue becomes more complicated in the case of heterogeneous structured populations (e.g. the populations of house-holds structured by the occupational status of the head of the house-hold, by income class, by number of children, etc., the populations of firms structured by number of employees, by activity, by profit volume etc.) its solving consisting in applying survey schemes with unequal probabilities.

References

Chelcea, S. (1995), *Cunoaşterea vieţii sociale. Fundamente metodologice*, Ed. I.N.I., Bucureşti.

Isaic-Maniu, Al., Vodă, V. (2009), „The Poisson property in the case of some continuous distributions", *Romanian Statistical Review*, vol. 2, pp. 56-64.

Isaic-Maniu, Al., Vodă, V. (2008), *Abordarea şase sixgma. Interpretări. Controverse. Proceduri*., Editura Economică, Bucureşti.

Isaic-Maniu, Al., Mitruţ, C., Voineagu, V. (2006), *Statistică*, Ed. Universitară, Bucureşti.

Isaic-Maniu, Al., Vodă, V. (2006), *Proiectarea statistică a experimentelor*, Editura Economică, Bucureşti.

Isaic-Maniu, Al. (coord.) (2003), *Dicţionar de statistică generală*, Ed. Economică, Bucureşti.

Isaic-Maniu, Al. (2001), *Tehnica sondajelor şi anchetelor*, Ed. Independenţa Economică, Bucureşti.

Isaic-Maniu Al., Mitruţ C., Voineagu V. (2000), *Statistica pentru managementul afacerilor*, (ed. a II-a), Ed. Economică, Bucureşti.

Mărginean, I. (2000), *Proiectarea cercetării sociologice*, Ed. Polirom, Iaşi.

Mihoc, Gh., Urseanu, V. (1985), *Sondaje şi estimaţii statistice*, Ed. Tehnică, Bucureşti.

Moser, C. A. (1967), *Metode de anchetă în investigarea fenomenelor sociale*, Ed. Ştiinţifică, Bucureşti.

Rotariu, Tr., Iluţ, P. (1997), *Ancheta sociologică şi sondajul de opinie*, Ed. Polirom, Iaşi.

Sandu, D. (1993) *Statistică în ştiinţele sociale*, Universitatea Bucureşti, Bucureşti.

Stoetzel, J., Girard, A. (1975), *Sondajele de opinie publică*, Ed. Ştiinţifică şi Enciclopedică, Bucureşti.

Trebici, V. (coord.) (1985), *Mică enciclopedie de statistică*, Ed. Ştiinţifică şi Enciclopedică, Bucureşti.

Yule, G., Kendall, M.C. (1969), *Introducere în teoria statisticii*, Ed. Ştiinţifică, Bucureşti.
Cochran, William G. (1977), *Sampling Techniques* (Third ed.), Wiley, New York.
Bartlett, J. E. II, Kotrlik, J.W., Higgins, C. (2001), „Organizational research: Determining appropriate sample size for survey research", *Information Technology, Learning, and Performance Journal*, Vol. 19, No. 1, pp. 43-50, accessed December 12th, 2009, from http://www.osra.org/itlpj/bartlettkotrlikhiggins.pdf
* * * Creative Research System, *Sample Size Calculator* accessed January 17th, 2010, from www.surveysystem.com/sscalc.htm