Definition of Sample Size of Heavy Metals Concentration in Various Soils

L.L. Frolova[1] and ***A.G. Zakirov***[2]

[1]Kazan State University, [2]Kazan Institute of Ecology of TAS, Russia
E-mail: Lucy.Frolova@ksu.ru

Keywords: evaluation data, software, confidence interval, sample

1 Introduction

The problem of the quantity and the quality of data in the processing of small sample size is actual in the various applications of biology, medicine and other scientific fields [1,5]. According to standard criteria and parametric statistical methods, the sample size must be large [3,4]. As a knows, stages of collection, input, storage, estimation of representation, primary statistical processing, definition of correlation between various parameters, etc., represent a technological chain of processing the data ecological. These stages are necessary for any usage of the initial data. Among these stages, the evaluation of the quantity and the quality of data collection should be considered as a key factor. The representative of the data is very important and the formal solution of this problem has a large practical interest. For these purposes the authors offer PC's program "Sample size"[2]. On the basis of the procedure proposed by C.Stein [6], the formula for definition the minimal size of sample with guarantered statistical stability of relevant conclusions, based on the given sample was deduced by authors. The two steps' procedure for definition of guarantee moment of a stop allows the researcher to determine the minimal sample size and based on construction of the confidence interval of fixed width for average normal distribution. It is possible to conduct, if process of a choice is carrying out in two stages, and the sample size is developing only at the second stage - casual size dependent on results of supervision at the first stage. The program "Sample size" is the part of database software "Servis-base" developed by authors.

2 Description of the model

The evaluation of the quantity and the quality of data collection are given below.
x_1, x_2, ... - sequence normal, equally distributed casual sizes $N(\mu,\sigma^2)$, where
μ, σ — unknown parameters:

$$-\infty < \mu < \infty$$

$$-\infty < \sigma < \infty$$

Show at first, that on the sample of fixed volume n it is impossible to construct an interval of fixed width for μ. For the sample of the volume n statistics $(\overline{x}_n, Q_n)$ - minimal sufficient statistics:

$$\overline{x_n} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

$$Q_n = \sum_{i=1}^{n} (x_i - \overline{x})^2$$

It is easy to prove, that there is no such statistics

$$P_{\mu,\sigma}\{|\,\mu - \hat{\mu}(\overline{x}_n, Q_n)\,| < \delta\} \quad \gamma$$

for all (μ, σ^2), where δ - is given, $0 < \delta < 1$, and $\gamma > 0$, respectively.

$$\sup_{\hat{\mu}} \inf_{\mu,\sigma} P_{\mu,\sigma} \{|\mu - \hat{\mu}(\bar{x}_n, Q_n)| \ > \delta\} \qquad \lim_{\sigma\to\infty} \inf_{\mu} \sup_{\mu} P_{\mu,\sigma} \{|\mu - \hat{\mu}\ (\bar{x}_n, Q_n)| < \delta\}$$

For the given meaning of parameter minimax statistics determining centre of an interval, is $\hat{\mu}(\bar{x}_n, Q_n)| = \bar{x}_n$, that is

$$\inf_{\mu} \sup_{\hat{\mu}} P_{\mu,\sigma} \{|\mu - \hat{\mu}(\bar{x}_n, Q_n)| < \delta\} = 2\Phi\left(\frac{\delta}{\sigma}\sqrt{n}\right) - 1$$

$$\to \sup_{\hat{\mu}} \inf_{\mu,\sigma} P_{\mu,\sigma} \{|\mu - \hat{\mu}(\bar{x}_n, Q_n)| < \delta\} < \lim_{\sigma\to\infty}\left[2\phi\left(\frac{\delta}{\sigma}\sqrt{n}\right) - 1\right] = \psi$$

That proves the statement, that there is no confidential interval of fixed width δ for μ in case of fixed sample size.

The following two-step selective procedure provides a confidential interval of fixed width δ for μ.

Step 1: undertakes sample of fixed volume n_0; $n_0 \geqslant 2$ and find of its sufficient statistics (x_{n0}, Q_{n0}).

Step 2: undertakes additional sample (casual) volume:

$$n_1 = \left(leastwholen, \quad \left[\frac{t^2 S_{n0}^2}{\delta^2} - n_1\right]^+ \right)$$

where $\quad a^+ = \max(\psi, a), S_{n0}^2 = Q_{n0} / (n_0 - 1)$

t – quantile 95% of student’s distribution sample size.

Notice, that n_1 – the sample size at the second stage - casual size, which can accept zero meaning with positive probability. If n_1=0, the process stops after the first stage. It is obvious, what $P\ \{n_1 < \infty\} = l$ for all (μ,σ).

Thus, n_1 – casual size. Besides it depends only from S^2_{n0} and does not depend from $\bar{x}_{n_0}$ and $(x_{n_0} + 1..., x_N)$ when $n_1 > l$. In this case average probability of covering by an average interval of fixed width is not less δ for all (μ,σ), $I_\delta(\bar{X}N)$ - confidential interval of fixed width with confidential level at γ.

Necessary volume of sample size can be found as:

$$v = \max\left(n_0, \left[\frac{t^2 S_{n0}^2}{\delta^2}\right] + 1\right)$$

where n_0 – first component of necessary volume of sample,

$\left[\frac{t^2 S_{n0}^2}{\delta^2}\right] + 1$ - second component, or guarantee moment of a stop.

The result of computation is representative the sample size for Confidence interval for Different in Means – 95 percent. The software “Sample size” was designed for realize this model.

3 Application of the model

In the test computations were used the data on the concentration of total and soluble forms of Pb in various soils for the area of Predvolzhye (Region of Middle Volga River) provided by the Institute of

Ecology of TAS. As a result of the program "Sample Size" application, the stable data on the sample size for the assigned half-width-confidence- interval (5%—30% of the mean value) have been computed. Table 1 shows the data on the concentration of total and soluble forms of Pb in various soils for the area of interest, obtained with the aid of PC's program "Sample Size", as well as results of analysis for normalized distributions for this data.

The program "Sample size" was tested for providing comparative computations on analysis of small sample size in order to check whether data can be approximated by normal distribution curves or not. For this purpose, the Shapiro-Wilks standard tests as well as the Kolmogorov-Smirnov criterion realized in the "Statgraphics Plus for Windows" software, were used. The last test is characterized by a more rapid convergence when applied in cases of the limiting distribution, i.e. they can be used for small sample size. The Kolmogorov-Smirnov test is generally regarded to be more powerful compared with the χ^2 criterion .

The assumption of the normalized distribution of the input data should be made for program "Sample size", if the calculated volume of representative data is equal to or less than that for the input data.

As follows from analysis of the data shown in Table 1, in most cases the sampling data obey the normal distribution law for the 20%—30% half-confidence-interval. Using the "Sample Size" program, the required sample size is found also. Moreover, the 20%—30% half-width-confidence-interval provides relatively low accuracy, which is, however, frequently typical for ecological investigations. As may be seen, the sampling procedure is quite sensitive to changes in the confidence interval. Notice, than less half-width-confidence interval than higher of the quality of data. Analyses for agreement of the conclusions made based on the "Sample Size" program, on one hand, and done with the use of the Shapiro-Wilks and Kolmogorov-Smirnov test, on the other hand, are represented in Tables 2 and 3.

Table 1 Results of analyses for normal distribution for concentrations of both total and soluble forms of Pb in various soils in some areas of the Volga Region

Type and subtype of soil	Metal	Expe-riment sample size	"Sample Size" Program						Tests for normality	
			5%	10%	15%	20%	25%	30%	Shapiro-Wilks	Kolmogorov-Smirnov
Brown-gray forest	Pbt	13	106	27	12	7	5	3	0.0611545 (+)	0.383143 (+)
	Pbs	13	322	81	36	21	13	9	0.504126 (+)	0.948208 (+)
Dark-gray forest	Pbt	8	96	24	11	6	4	3	0.656056 (+)	0.999997 (+)
	Pbs	10	401	101	45	26	17	12	0.100488 (+)	0.836255 (+)
Chernozem	Pbt	8	44	11	5	3	2	2	0.60509 (+)	0.994132 (+)
	Pbs	9	234	59	26	15	10	7	0.0820732 (+)	0.547587 (+)
Turf-podzol	Pbt	6	104	26	12	7	5	3	0.0283889(-)	0.727381 (+)
	Pbs	7	224	56	25	14	9	7	0.680225 (+)	0.971808 (+)
Chernozem -podzol	Pbt	19	59	15	7	4	3	2	0.272127 (+)	0.743984 (+)
	Pbs	21	757	190	85	48	31	22	0.0000057 (-)	0.181074 (+)
Chernozem - alkaline	Pbt	22	45	12	5	3	2	2	0.634036 (+)	0.97063 (+)
	Pbs	22	327	82	37	21	14	10	0.0062516 (-)	0.543918 (+)
Light- gray forest	Pbt	32	99	25	11	7	4	3	0.0535257 (+)	0.477602 (+)
	Pbs	33	529	133	59	34	22	15	2.0306E-7 (-)	0.0061222 (-)
Gray forest	Pbt	33	65	17	8	5	3	2	0.0005266 (+)	0.45837 (+)
	Pbs	37	559	140	63	35	23	16	0.0008975 (-)	0.140217 (+)

As follows from comparative analyses of the data in these tables, for the 30% half- width-confidence-interval, the conclusions from the "Sample Size" procedure in 81 % of cases agree with the ones made on the basis of the Kolmogorov-Smirnov test; and in 69 % of cases coincide with the ones involving the Shapiro-Wilks criterion. In 75 % of cases, the conclusions made by the above criteria are in agreement (Table 4). So, the test are given by program "Sample size" has a high result.

In all tables the sign "+" corresponds to the coincidence of conclusions made by the particular test and "Sample Size" program, including cases when, for example, the particular test rejects for normality where as the "Sample Size" program provides the conclusion that the sample size is not representative; the sign "–" indicates that there is no agreement of conclusions made by the standard tests and "Sample Size" program, $^{*)}$ Pbt = total Pb; Pbs = soluble Pb.

Table 2 Analysis for coincidence of the conclusions obtained by Shapiro-Wilks test and "Sample Size" program

Type and subtype of soil	Metal	10%	15%	20%	25%	30%
Brown-gray forest	Pbt	-	+	+	+	+
	Pbs	-	-	-	+	+
Dark-gray forest	Pbt	-	-	+	+	+
	Pbs	-	-	-	-	-
Chernozem	Pbt	-	+	+	+	+
	Pbs	-	-	-	-	+
Turf-podzol	Pbt	+	+	+	-	-
	Pbs	-	-	-	-	+
Chernozem -podzol	Pbt	+	+	+	+	+
	Pbs	+	+	+	+	+
Chernozem - alkaline	Pbt	+	+	+	+	+
	Pbs	+	+	-	-	-
Light- gray forest	Pbt	+	+	+	+	+
	Pbs	-	+	+	-	-
Gray forest	Pbt	+	+	+	+	+
	Pbs	+	+	-	-	-
Coincidence	quantity	8	11	10	9	11
	percent	50	68.8	62.5	46.3	68.8

Table 3 Analysis for coincidence of the conclusions obtained by Kolmogorov-Smirnov test and "Sample Size" program

Type and subtype of soil	Metal	10%	15%	20%	25%	30%
Brown-gray forest	Pbv	+	+	+	+	+
	Pbp	-	-	-	+	+
Dark-gray forest	Pbv	-	-	+	+	+
	Pbp	-	-	-	-	-
Chernozem	Pbv	-	+	+	+	+
	Pbp	-	-	-	-	+
Turf-podzol	Pbv	-	-	-	+	+
	Pbp	-	-	-	-	+
Chernozem -podzol	Pbv	+	+	+	+	+
	Pbp	-	-	-	-	-
Chernozem - alkaline	Pbv	+	+	+	+	+
	Pbp	-	-	+	+	+
Light- gray forest	Pbv	+	+	+	+	+
	Pbp	-	-	-	+	-
Gray forest	Pbv	+	+	+	+	+
	Pbp	-	-	+	+	+
Coincidence	quantity	5	6	9	12	13
	percent	31.3	37.5	46.3	75	81

Table 4 Agreement of conclusions made by the Shapiro-Wilks and Kolmogorov-Smirnov tests

Type and subtype of soil	Metal[*)]	Coincidence of test conclusions
Brown-gray forest	Pbt	+
	Pbs	+
Dark-gray forest	Pbt	+
	Pbs	+
Chernozem	Pbt	+
	Pbs	+
Turf-podzol	Pbt	-
	Pbs	+
Chernozem -podzol	Pbt	+
	Pbs	-
Chernozem - alkaline	Pbt	+
	Pbs	-
Light- gray forest	Pbt	+
	Pbs	+
Gray forest	Pbt	+
	Pbs	-
Coincidence	quantity	12
	percent	75

4 Conclusion

This approach gives a useful preliminary information about a possibility of the parametrical description of an available data set and the sample size necessary for deriving of stable conclusions with the further statistical processing.

Given statistical approach to the evaluation of the quantity and quality of data collection allows to plan the list of heavy metals and the quantity of samples which necessary to have in a future during a new field's researches.

References

[1] Aivazyan, S.A., Enukov, I.S. and Meshalkin, L.D.: 1983, *Fundamentals of Modeling and Primary Data Processing* (in Russian). Finance and Statistics Publisher, Moscow, 471p.

[2] Frolova, L.L., Zakirov, A.G. and Koroleva, T.E.: 1994, *The Evaluation of Data Representation.* In: Proceedings of the Second Conference on Advanced Biochemical Engineering, Brighton, UK, pp 160-162.

[3] Lloid, E. and Lederman, W. (eds): 1989, *A Handbook on Applied Statistics* (in Russian), v.**1** and **2**. Finance and Statistics Publisher, Moscow.

[4] Orlov, A.I.: 1995. To the Agreement Criteria for the Parametric Family (in Russian). *Factory Laboratory*, v.**61** [7], pp.59-61.

[5] Petrov, A.A.: 1956, Examination of Statistical Hypotheses on a Distribution Type for Small Samples (in Russian). In: *Theory of Probability and Its Applications*, v.**1** and **2**, pp. 248-269.

[6] Zaks, S.: 1975, *Theory of Statistical Derivations* (in Russian). Mir Publisher, Moscow, 776p.