4/1 (2006), 89–99
tmcs@inf.unideb.hu
http://tmcs.math.klte.hu

Teaching
Mathematics and
Computer Science

Understanding the spatiotemporal sample: a practical view for teaching geologist students

Ilona Kovácsné Székely

Abstract. One of the most fundamental concept of statistics is the (random) sample. Our experience – acquired during the years of undergraduate education – showed that prior to industrial practice, the students in geology (and, most probably, in many other non-mathematics oriented disciplines as well) are often confused by the possible multiple interpretation of the sample. The confusion increases even further, when samples from stationary temporal, spatial or spatio-temporal phenomena are considered. Our goal in the present paper is to give a viable alternative to this overly mathematical approach, which is proven to be far too demanding for geologist students.

Using the results of an environmental pollution analysis we tried to show the notion of the spatiotemporal sample and some of its basic characteristics. On the basis of these considerations we give the definition of the spatiotemporal sample in order to be satisfactory from both the theoretical and the practical points of view.

Key words and phrases: spatiotemporal sample, random phenomenon, non-mathematics students.

ZDM Subject Classification: K10, K40.

1. Introduction

Comprehension and proper application of the basic notions and methods of statistics assumes good establishment in probability theory. One of the most fundamental concepts of statistics is the (random) sample [1, 3].

In the industrial practice the "sample" is neither defined uniquely nor in full extent. In many cases only one element of the mathematical sample is regarded

as "the sample". On the contrary, sometimes the mean of certain parameters measured or analyzed at the same time and location under the same circumstances may also be called "sample". E.g. in geology, an averaged quality parameter, weighted by the thickness at a certain location, given by its x and y coordinates, constitutes a "sample" [2, 4].

Our experience during the undergraduate education showed that prior to industrial practice, the students in geology (a most probably in many other non-mathematics oriented disciplines as well) are often confused by the possible multiple interpretation of the sample. The reason is that when a real life (e.g. geochemical) application is given, the student encounter a realization of the sample, and they often identify it with the theoretical sample itself. The descriptive statistics (mean and variance most commonly) computed from the sample appear for them as single parameters, or single values and do not understand why should they be regarded as random variables. As in most cases the replication of the sample is not possible, or at least not in a straightforward way, they ignore the fact that the realisations vary, and so will vary the statistics computed from the realisations. The confusion increases even further, when samples from stationary temporal, spatial or spatio-temporal phenomena are considered.

In this case the student encounters e.g. the averaging of measurements taken only at one occasion but at different times and/or location, while he/she learnt up to now, that the statistics summarises the information of samples taken under identical and independent conditions. The ergodic theorem dissolves the virtual contradiction, but how should it be explained to the geologist students. A way how shouldn't is for sure: forget about introducing the ergodic theorem! Our goal in the present paper is to give a viable alternative to this mainly mathematical approach, which is proven to be far too demanding for geologist students.

2. Theoretical definition of the sample

In the simplest and most common situation a random phenomenon is characterized by a certain random variable X with known distribution, and one can observe its values at n independent occasions under theoretically identical conditions. As it is often said an independent sample of size n is drawn from the distribution of X.

So, the *independent statistical sample* from a one dimensional distribution Q_X consists of n independent identically distributed random variables denoted by $(X_1, X_2, \ldots, X_n)$ whose common distribution is Q_X. This applies to the mean

and the variance (the existence of which we suppose all throughout this paper), too: $E(X_i) = E(X) = m$; $D^2(X_i) = D^2(X) = \sigma^2$ where $i = 1, 2, \ldots, n$.

The *realisation of the (independent statistical) sample* consists of n real numbers $(x_1, x_2, \ldots, x_n)$, the actually observed values of those random variables: $X_1 = x_1$, $X_2 = x_2$, $\ldots$, $X_n = x_n$.

This formulation, based on the duality of random variables and numbers, is rather difficult for non-mathematics students to understand. And this is the point where the confusion increases, when one turns to temporal, spatial or spatio-temporal phenomena.

The full analogue of the above definition of i.i.d. sample for the case of a random phenomenon dependent on location and time characterised by a random spatio-temporal process $Z(s, t)$, would require several independent measurements at every single location and time-point. But more often than not we have only one chance to measure. It is a nonsense to measure several times independently e.g. the aluminium oxide concentration at a certain location in a mine or the water discharge of a river at a gauge station at a given time. It is the time-shift or the location-shift invariance, that enables us to use the information collected at other sites or at other times. And this is what the student has to understand well, and can serve as the starting point for understanding the sampling of a stationary phenomenon.

Having a real valued spatially and temporally changing random process, it can be decomposed theoretically into random variables on an infinite number of elementary units having zero volume in the original state space by restricting it to a fixed space-time point. Consequently in the investigation of certain variable (parameter in the geological terminology) of the phenomenon we obtain, in principle, a sample realisation of infinite size. Under the assumption of stationarity (or more strictly, ergodicity) this sample realisation will be a realisation of identically distributed random variables. Disregarding of the trivial case of an independent value noise, however, independence will not be satisfied for this sample. This data set can then be regarded as the one to be analysed. So, the realisation is informative on the statistical characters of the distribution, like the expectation and the variance, and the mean will converge to the theoretical value of the expectation, however, because of the correlated samples, the usual statistical properties of the corrected empirical variance estimator does not remain valid in full extent. We illustrate the infinite manifold on Figure 1. In this figure we divided the "area" of the observation domain into "infinite" number of elementary units of zero volume.

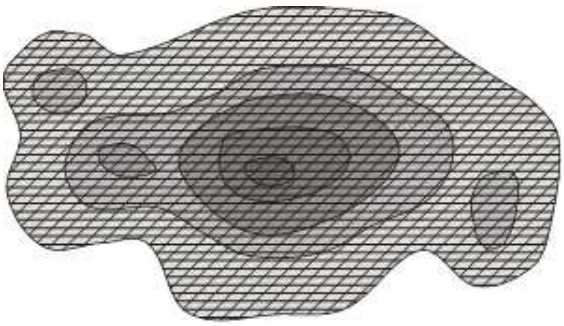


Figure 1. Isoline figure of a "parameter" of a theoretical phenomenon.

Mainly because of financial and/or technical reasons, in concrete projects there is seldom a chance for taking a really large (eventually thousands of) sampling, that would provide great accuracy in estimating the statistical characteristics. In addition, in practice there appears another problem as well, namely that the volume of the elementary units is not zero but greater. In the majority of cases there is no chance to take a nearly infinite number of samples, or to repeat the samples several times for a smaller size. So, the situation is that $V \gg 0$, but compared to the magnitude of the phenomenon $V \approx 0$, that is the variability within the elementary units is ignorable compared to the variability on the larger scale, but $N \ll \infty$. This raises the question, whether a sample of this kind can be regarded as representative, does it really reflects the property of the manifold it has been taken from. Figure 2 illustrates a possible sample realisation for $N < \infty$ and $V > 0$.

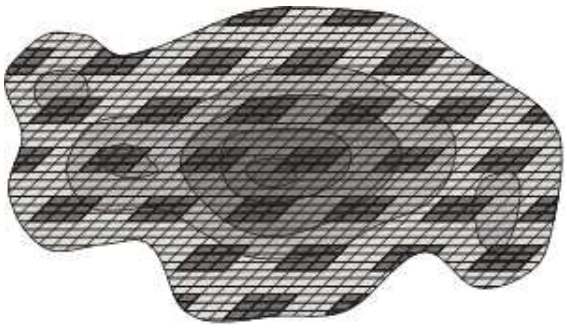


Figure 2. A possible sample realisation for $N < \infty$ and $V > 0$.

The most important property of a variable (geological parameter) of a phenomenon – supposing uniform placement and sufficient number of samples – the probability distribution. **In practice the type of the distribution is not known, its identification is an important task.** According to experience at least a sample size of 50 is needed in order to draw conclusions on the type of distribution from the histogram of the variable. On the basis of the sample an estimation can be given for the main characteristics of the sample, e.g. for the expectation and the variance.

3. The notion of sample and some of its properties illustrated in an example

In practice the need for creating a sufficiently large spatiotemporal sample that can fulfil the role of the basic example of a manifold under study, encounters almost unresolvable financial difficulties. Therefore it is a great advantage, that the dataset of measured chemical components of an infamous really-happened environmental pollution can be used for clarifying the notion of the sample, the computation of its parameters and the identification of its distribution.

From the start of the 1900's in the 22nd district of Budapest, in the densely populated Nagytétény area a factory worked as a constant source for pollution emitting harmful substances into the air. Considerable heavy metal pollution befell the area, and by the cultivation of the land it penetrated into the deeper layers of the soil, too. An environment-protectional fact-finding action was carried out by soil analysis, taking samples on the 3.5 km^2 private properties and land by 20 cm layers. This means more than 1000 soil sample for the uppermost 0–20 cm layer to be analysed for 24 chemical elements or components. We choose 3 of this 24: calcium, phosphor, and arsenic that are random variables in the present model. The measurement unit is mg/kg for every element.

We consider the measurement results for chemical components as elements of the statistical manifold. In our opinion the 1026 and 1100 sample size in the given case is large enough to approximate an infinite sample and it is appropriate to illustrate the tracted statistical concepts, without breaking the requirements of the mathematical theory considerably.

However, as we obtained a finite manifold, the variances and means of the random variables became computable. They are given in Table 1.

Table 1. The parameters of the manifold.

	Sample size	Mean	Std. deviation
Ca	1026	71429.79	25301.74
As	1110	17.16	23.33
P	1026	1652.86	1014.58

In order to demonstrate that the individual sample elements are random variables, we took random 100 long resampled observations 1000 times from the manifold of Ca, As and P variables. Table 2 shows parts from the realisations of Ca-samples. It is clear that the realisations of the 100 long samples vary from sample to sample.

Table 2. Realisations of Ca-samples.

Calcium	Realisations of samples						
	X_1	X_2	X_3	X_4	...	X_{99}	X_{100}
1. sample	58339.44	52771.68	59438.88	44729.46	...	80643.21	73601.96
2. sample	78266.18	82664.06	59843.29	61782.66	...	55465.67	47424.37
⋮							
1000. sample	51623.78	59682.54	45447.23	50109.89	...	8806.80	8452.38

Descriptive statistics can be computed from the samples of Table 2, of which the most important, the mean is presented in Table 3. We extended Table 3 for the other two chemical components As and P. The values displayed in the table present illustratively the statement, that the mean is also a random variable, it changes from sample to sample and its values vary around the grand mean, that is the mean of the complete original sample (Table 1).

Table 3. Realisations of the sample means.

	Sample mean		
	As	Ca	P
1. sample	20.94	66420.70	1741.25
2. sample	15.77	69815.47	1775.85
⋮			
1000. sample	14.36	74512.84	1542.36

The average of all possible sample means gives the mean of the manifold, that is the expected value. This property means the unbiasedness of the estimation $E(\overline{X}) = m$. Could we take into account all possible sample means, only then it would be possible to demonstrate this theoretical statement. However, this leads to enormous difficulties, because e.g. in the case of Ca, from the 1026 element that we regarded as manifold, it is possible to choose a 100 long subsample in $9.51 \cdot 10^{140}$ different ways. In other words we can resample from this manifold this many 100 long realisations, consequently this many means can be computed. To carry it out in practice is simply impossible. (There are $1.288 \cdot 10^{307}$ ways to select a 500 long subsample.) It is possible to demonstrate only that the sample means approximate the grand mean (the mean of the complete manifold) well, and the error of the approximation is decreasing as the sample size increases. The results presented in Table 4 demonstrate the validity of our statement. We carried out this program by taking 100, 300, 500 long subsamples 1000 times from the manifold, and computing the averages and the standard error of the sample means. Consider the distribution of the sample means of the Ca, As and P variables as in Table 4. The histograms in Figures 3(a), 4(a), 5(a) show the distribution of the complete Ca, As and P manifolds. The empirical probability density estimations of the sample averages for 1000–1000 subsamples, 100, 300 and 500 long each, are displayed in Figures 3(b), 4(b), 5(b); 3(c), 4(c), 5(c); 3(d), 4(d), 5(d) respectively.

Table 4. Averages and standard errors of the sample means.

Random variable – sample realisation	Sample size	Average of sample means	Standard error of sample means
Ca-100	1000	71498.90	74.72
Ca-300	1000	71399.45	39.01
Ca-500	1000	71444.89	25.12
As-100	1000	17.26	0.07
As-300	1000	17.12	0.04
As-500	1000	17.16	0.03
P-100	1000	1658.98	2.98
P-300	1000	1652.99	1.60
P-500	1000	1652.09	1.00

It can be seen very clearly in all cases, that the sample mean follows a normal probability law independently from the distribution of the original manifold. The distributions get closer to normal as the sample size increases.

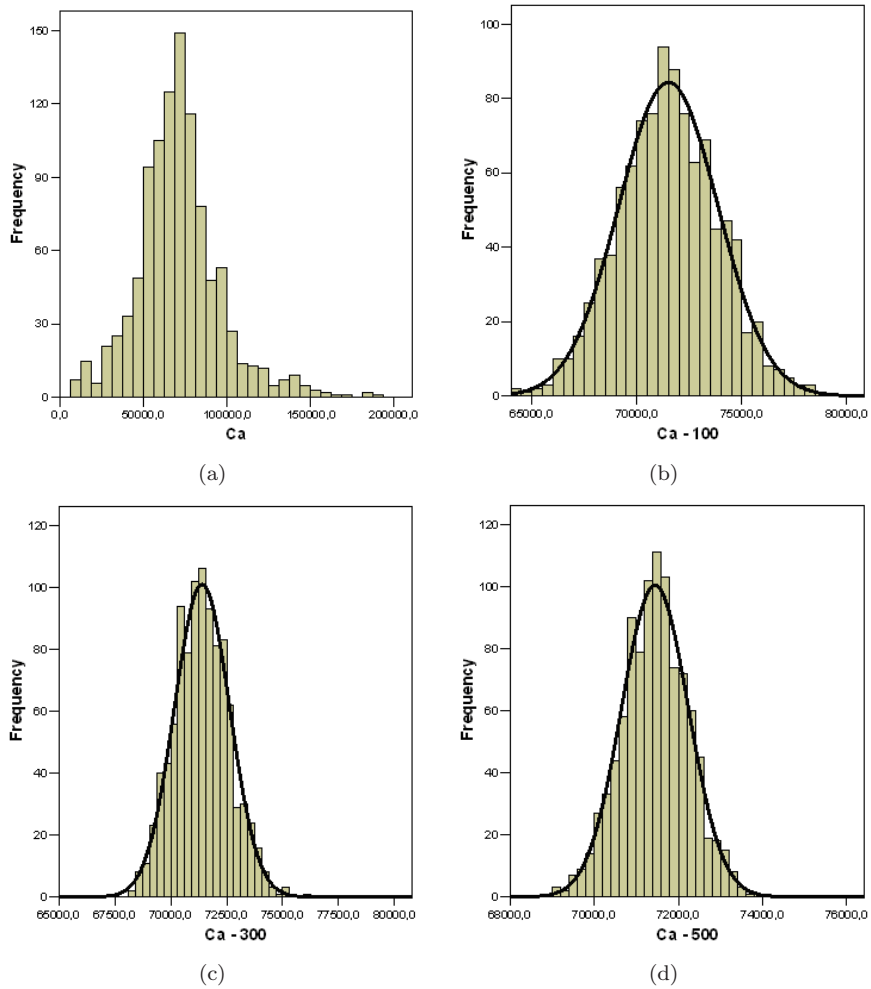


Figure 3

The observations of Ca were regarded in the preceedings as 1026 samples. In fact we treated these observations as if they were independent identically distributed ones. However, an other approach is also viable. We can also consider

the spatial dependence structure. In this case we have only one sample of 1026 dimensions. It is worth to display it as a map, like in Figure 2. This approach requires the use of various spatial models, but detailed elaboration on this topic goes beyond the framework of the present paper.

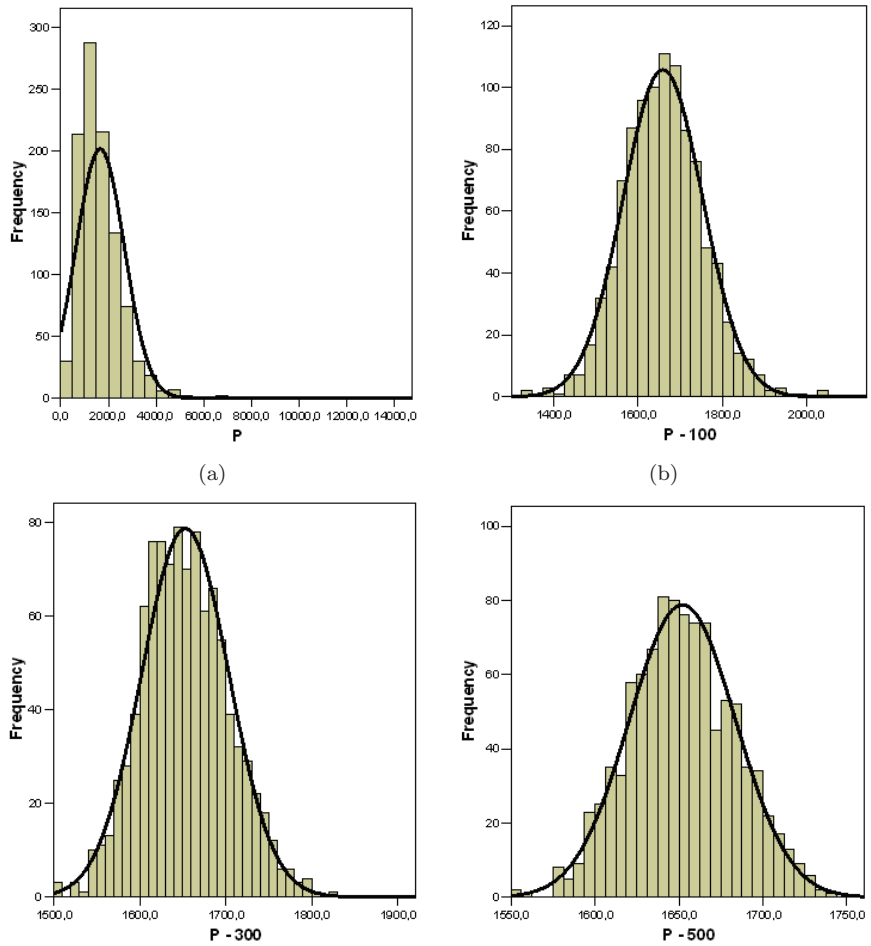
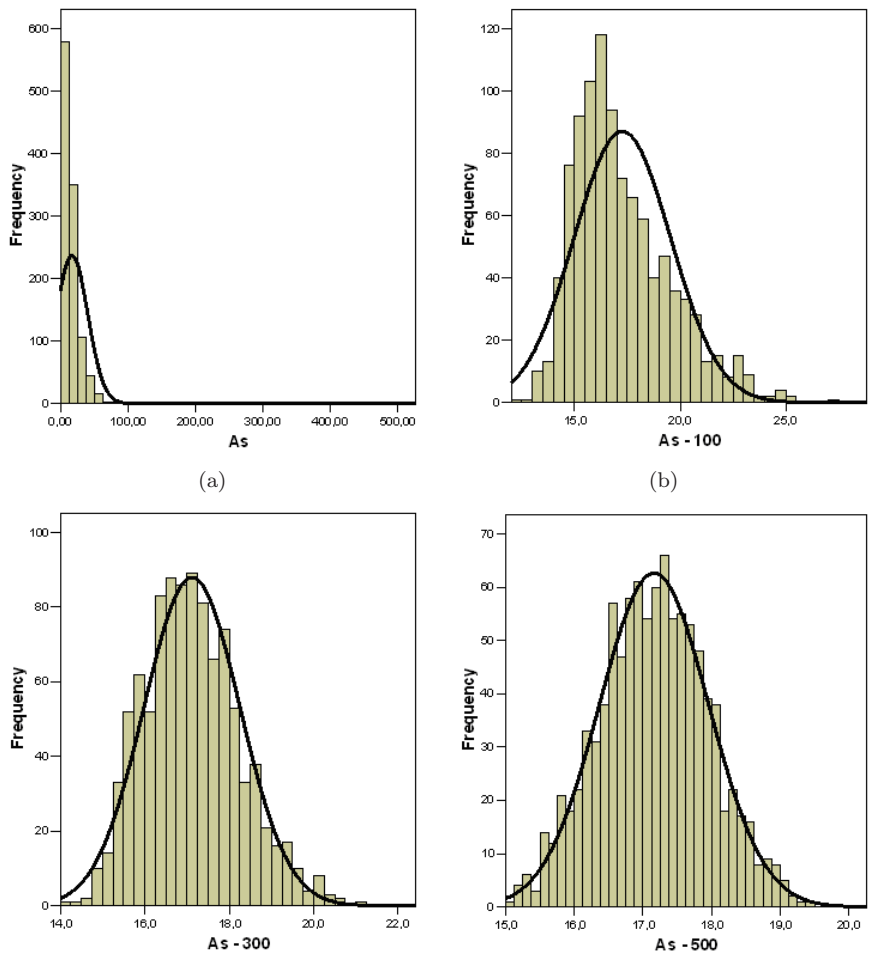


Figure 4



(a) (b) (c) (d)

Figure 5

4. Conclusion

Using the results of an environmental pollution analysis we tried to show the notion of the spatiotemporal sample and some of its basic characteristics. On the basis of these considerations we give the definition of the spatiotemporal sample

in order to be satisfactory from both the theoretical and the practical points of view. We propose the following definition:

In practical sense the values of a parameter of certain phenomenon that can be associated with x, y, z, t coordinates and either measured in situ, analysed or computed are called a spatiotemporal sample. The practice doesn't use the adjective spatiotemporal, but we should here, in order to distinguish the present considerations from the classical theory. *The sample in the practical sense corresponds to one element of the mathematical sample, with the difference that it is associated with a space-time unit with non-zero volume.*

References

[1] P. W. Anderson, R. M. Loynes, *The Teaching of Practical Statistics*, John Wiley & Sons, Chichester, New York, Brisbane, Toronto, Singapore, 1987.

[2] A. Füst, *Geostatistics*, Eötvös Kiadó, Budapest, 1998 (in Hungarian).

[3] T. Nemetz, G. Wintsche, *Probability and Statistics for students*, Polygon, Szeged, 1999 (in Hungarian).

[4] R. Webster, M. A. Oliver, *Statistical Methods in Soil and Land Resource Survey*, Oxford University Press, 1990.

ILONA KOVÁCSNÉ SZÉKELY
FELVINCI ÚT 15.
H–1022 BUDAPEST
HUNGARY

E-mail: iszekely@geology.elte.hu

(Received August, 2005)