

How Many Samples Are Enough?

Introduction

When Cliff Stanley telephoned me last April and asked me where the '30 samples' came from, I had to admit that I thought I was responsible and that the estimate came from a consideration of the standard error of the mean. I could not remember at the time where I had written it and illustrated the plot of $1/\sqrt{n}$ versus n ; it was not in my first guess, the chapter on sampling in Richard Howarth's Volume 2 of the Handbook of Exploration Geochemistry (Garrett 1983a). Some digging retrieved the 1979 paper (Garrett 1979) that I forwarded to Cliff, and set me thinking as to just why I chose 30.

The choice was not quite as arbitrary as it may appear, what was arbitrary was another number, 120. In the days of statistical tables, the Student-t statistic, used for estimating confidence bounds on means, was only tabled up to 30 (see for example, Mathematical Tables from Handbook of Chemistry and Physics, 11th Edition, 1963), or occasionally 120, as after that it changed only very slowly and slightly up to the value for an infinite sample size. Today the required values of statistics are computed 'on the fly' by the software in the PCs we all use, as are the probabilities associated with hypothesis tests based on those statistics. With reference to the value of $1/\sqrt{n}$, by the time n reaches 30 the value of $1/\sqrt{n}$ has fallen to 90% of that which it attains at 120; and the first derivative, i.e. the rate of change for the function, had fallen to well in excess of 99% of its value for 120.

The choice of the round number 30 did not seem too unreasonable if the objective was to estimate a mean for what was thought to be a homogeneous terrane, rather than set out a network of samples to detect a dispersion halo from a mineral occurrence. In this latter instance a completely different set of guidelines are applicable; primarily to ensure the area of the sampling grid cells is smaller than the area of geochemical dispersion pattern being sought, preferably by a factor of two or more, and that the grid is favourably oriented relative to the expected orientation of the dispersion patterns. Financial costs are always a major concern in designing and executing geochemical surveys, for estimating means, things were not going to change too radically after the sample size reached 30. Unless, of course, the unit or area being sampled was not, in fact, homogeneous, i.e. the issue of 'a specific situation'.

How many samples really are necessary?

We have to thank Cliff for returning to this question in his Explore article (Stanley 2010) and proposing that it needs to be asked within the context of a particular survey in a relative way. Not long before I received his phone call I had been asked the same question in a workshop in Halifax, NS, by an environmental consultant concerned with undertaking risk assessments. I answered in the traditional manner, "30", and was horrified to hear that he thought 3 were sufficient in order to characterize soils in his study areas to compare with the CCME (1997) guideline for As in soils. Granted the size of the areas the consultant was working with were small land parcels, not geological-scale lithozones. However, this in itself can cause problems because of spatial correlation

effects, as noted below in the concluding remarks.

Stanley (2010) points out that his derivation of the formula:

$$n \approx \left\{ \frac{CV}{rse_x} \right\}^2$$

is distribution free. True, however, that is not the case for the estimate of CV for the target population being sampled. Estimates of the classical mean and standard deviation (variance) are made with the underlying assumption that the data are normally and independently distributed. Some practitioners rely on the fact that the mean and variance are the most statistically efficient estimators, however, that is only true if the assumptions are met. In most cases regional geochemical data are neither normally distributed, even after some appropriate transformation, e.g., logarithmic, or square root if the underlying true distribution is Poisson, nor independent of one another due to spatial correlation effects.

It is now widely recognized that the estimation of the properties and range of geochemical background benefits from the use of robust estimators, see, for example, Reimann et al. (2005). This recognition is also exemplified by the wide use of the median by practicing geochemists rather than the mean as a measure of central tendency in regional surveys, though it must never be forgotten that background is a range and cannot be suitably defined by a single number (Reimann and Garrett 2005). In general, the adoption of robust estimators by applied geochemists has been slow, although they have been in use for more than 30 years (Garrett 1983b), and physical scientists have been urged to use them by statisticians, see, for example, Rocke et al. (1983). More recently the use of both robust estimators and non-parametric methods, which are resistant to the effects of outliers and distributional assumptions, has become more widespread and tools provided to assist the applied geochemist in their application (Reimann et al. 2008).

The practical problem described by Stanley in Nova Scotia, Canada, an outcome of the North American Soil Geochemical Landscapes project, will occur again. The project was proposed in 2001, and following several workshops a continent-spanning orientation survey was undertaken in 2004 (Smith et al. 2009). Following that, the Canadian and U.S. Geological Surveys commenced full scale sampling in 2007. To address the problem of having a sufficient number of samples to support the estimation of background statistics on a provincial-scale, sampling of the Maritime Provinces was undertaken at a double density of 1 site per 800 km², i.e. two 20 x 20 km cells sampled in every 40 x 40 km cell. The base sampling density of 1 site per 1600 km² led to an anticipated suite of approximately 6000 samples across North America. The Geological Survey of Canada withdrew from the tri-national effort in 2009, the U.S. and Mexican surveys are continuing.

Recognizing the high variability of geochemical data and the presence of outliers, for example, the 345.7 ppm As value in the 'Over Meguma Supergroup' data discussed by Cliff (Stanley 2010, Table 1), how can estimates of "how many samples are enough?" be improved. The specific value of

continued on page 10

How Many Samples Are Enough? continued from page 9

345.7 ppm As is probably not ‘stable’, a second soil sample collected a few meters away would likely vary by as much as 50%. The mean, standard deviation and coefficient of variation for the As data are 46.6, 84.8 and 181%, respectively. The act of cutting the high value of 345.7 by a half to 172.8 results in estimates of 35.1, 42.3 and 121%, for the same estimators respectively. The use of robust estimation procedures leads to estimators for the median, Median Absolute Deviation (MAD) and an Inter-Quartile Range (IQR) based estimate of the standard deviation (SD), see Table 1 below, that do not differ between the two data sets. Depending on which of the two analyses was used different estimates of “how many samples are enough?” would be obtained because of the sensitivity of variance and standard deviation estimates to outliers and extreme values. However, if robust estimation procedures were used there would be no difference in the estimates of required sample size, for this example, because robust outlier resistant procedures ‘down-weight’, reduce the effect of, outliers and extreme values.

Table 1 presents the classical statistical estimators from Stanley’s Table 1 and additional robust estimators, medians, and MADs and IQR based estimates for standard deviation (data spreads). Modern statistical packages compute medians and MADs, and an IQR based estimate of the standard deviation (SD) can be obtained by multiplying the IQR, the data range between the first and third quartiles (Q1 and Q3, or 25th and 75th percentiles) by 0.7413, the ratio between the 50% of the data encompassed by the IQR and the data

encompassed between mean $\pm$ 1 SD for a normal distribution. This leads to the estimation of robust CVs (RCVs).

Table 1.
Estimation of various coefficients of variation for As (ppm) concentration in soils.

Statistic	Over Granite	Over Meguma
Mean	13.35	46.61
SD	7.36	84.77
CV (%)	55.12	181.15
Median	13.80	22.20
MAD	4.45	13.05
RCV _{MAD} (%)	32.23	58.77
Q1	10.8	14.6
Q3	15.9	33.6
IQR	5.1	19.0
SD _{IQR}	3.88	14.46
RCV _{IQR} (%)	28.14	65.16

The new robust estimates may then be used to estimate “how many samples should be enough?”. Table 2 presents the results for a 10% tolerance level, as used in Cliff’s example.

Table 2.
Number of soil samples required to achieve a 10% tolerance level

	CV _{Classical}	CV _{MAD}	CV _{IQR}
Over Granite	30.38	10.39	7.92
Over Meguma	328.15	34.54	42.46

Conservatively, these would be rounded to 31, 11, 8 for ‘Over Granite’ soils, and 329, 35, 43 for the normal assumption for the ‘Over Meguma’ soils, respectively. With an appreciation of the natural variability of the geological environment I would not argue that fewer than 30 soil samples should be collected over the South Mountain batholith, it is known to be multi-phase. Furthermore, going back to the original 1979 reasoning, fewer than 30 samples and the confidence bounds around the mean may not be sufficiently tight for the purpose of the survey. However, I would certainly argue against collecting 330 soil samples from the Meguma Supergroup, perhaps 40 would be adequate. The choice ultimately belongs to the geochemist designing the survey, and the financial resources and time constraints for the project.

However, there is a more fundamental question that needs to be addressed, and that is the nature of the statistical distribution of As values in the Meguma Supergroup rocks and their overlying soils. The key issue is whether one is dealing with background data from a normal-like distribution contaminated by outliers, or data from a lognormal distribution that encompasses extreme values. In the former case, normal-like background data have been contaminated by

Note: This EXPLORE article has been extracted from the original EXPLORE Newsletter. Therefore, page numbers may not be continuous and any advertisement has been masked.

How Many Samples Are Enough? continued from page 10

outliers from an epigenetic As-Au mineralization process that led to the formation of the gold deposits that have been mined in the Meguma host rocks. In the latter case, the 345.7 ppm As value is simply an extreme value of a continuum of values that vary from very low all the way to extreme values, possibly even greater than 345.7 ppm As. The decision as to the form of the underlying statistical distribution has major implications. For the situation where contamination by an epigenetic process is suspected a robust estimation procedure should be used, not a classical procedure, the results of which are demonstrated in Tables 1 and 2 above. On this basis the calculation would indicate that between 35 and 43 samples should be adequate, assuming the same robust procedure would be applied to the resulting data for presentation, discussion and decision taking. If, however, a lognormal distribution is assumed for the Meguma Supergroup soils, and all relevant calculations are undertaken in logarithms, it would be appropriate to use 'classical' estimators because extreme values are to be expected in a lognormal, or any other strongly positively skewed, statistical distribution.

Table 3 presents the calculations for the soils overlying the Meguma Supergroup, robust estimates are not made as they are inappropriate if one accepts the hypothesis that the data are drawn from a lognormal, or approximately, lognormal distribution.

Table 3.

Summary statistics for the Meguma Supergroup soils assuming a lognormal distribution. Data Log base10 transformed, antilogged values in brackets.

	Mean (Geom. Mean)	Std. Deviation (SD)	CV%
As (ppm) in soil	1.392 (24.66)	0.4218	30.30

This estimate can now be used to estimate the number of samples that might be adequate under the assumption of lognormality. For the normal (arithmetic) examples a 10% tolerance was assumed. For the critical level of the CCME Soil Quality Guideline of 12 ppm, where decision taking is most critical, that results in a ± 1.2 ppm tolerance band. However, for a lognormal distribution the tolerance band converted back to the arithmetic scale the bounds are asymmetric. Thus a 10% tolerance at 12 ppm translates to 9.4 to 15.4 ppm; these could be considered too wide, whereas a 5% tolerance results in bounds of 10.6 and 13.6 ppm. On this basis it would be more appropriate to estimate required sample size for a 5% tolerance level for comparison between the two distributional models. The addition of the result for a 5% tolerance bound for the lognormal model is shown with the classical and robust estimates in Table 4.

From Table 4 it can be argued that a sample size of somewhere between 35 and 43 should be adequate to estimate the statistical properties of the soils overlying the Meguma Supergroup rocks and achieve suitable tolerance levels at the decision-critical CCME Soil Quality Guideline level of 12 ppm As. The selection of a sample size of 40 would satisfy both models, 'epigenetically' contaminated normal-like and lognormal for the As distribution, and lead to reasonable

confidence bounds on the estimated mean/median and geometric mean.

Table 4.

Number of soil samples required to achieve a 10% tolerance level for the assumption of a contaminated normal distribution, and a 5% tolerance level for the assumption of lognormality

	CV _{Classical}	CV _{MAD}	CV _{IQR}
Over Meguma (Normal)	328.15	34.54	42.46
Over Meguma (Lognormal)	36.72		

My own bias is to assume that the 'Over Meguma Supergroup' data are drawn from an approximately normal background population that has been contaminated by As from an epigenetic process associated with the formation of the Au-As vein systems in the Meguma host rocks. For that reason I believe the use of a robust statistical procedure is desirable for estimating population statistics, with the median and MAD being the obvious choices.

I argue that Stanley's new procedure can be improved by the application of robust estimators that would better focus on the properties of the background distribution, rather than base a decision on a data set contaminated by outliers from geochemical processes other than background. Clearly the 'outlier' in the sample size estimates, assuming a contaminated normal-like background distribution (Table 2), is that for classical estimation procedures applied to soils overlying Meguma Supergroup rocks. This demonstrates the vulnerability of decisions based on classical statistical procedures when dealing with the reality of natural processes and the presence, in this case, of just one outlier. As Rocke et al. (1982) asked, "Are robust estimators really necessary?", I would answer a loud, "Yes", and think of the amount of time and financial resources that would be saved.

Parting comments

Firstly, the coefficient of variation is a very useful summary statistic, expressing the variability of a data set relative to its central tendency, median or mean, see, for example, the discussion by Koch and Link (1971). During the early days of the National Geochemical Reconnaissance (ca. 1980) it was noted that high non-robustly estimated, i.e. mean and standard deviation based, CVs, >70%, indicated the presence of diverse geochemistry reflecting complex geology and/or the presence of mineral occurrences as the estimates were strongly influenced by positively skewed distributions and outliers. As such, these CVs were useful screening tools for geochemical data subdivided by source lithological unit or some other criterion, identifying areas that deserved additional interpretive attention (Garrett et al. 1980).

Lastly, geochemical data are spatially correlated, samples collected closer together tend to be more geochemically similar than those collected farther apart. Spatial correlation issues fall within the study of regionalized variables and the practice of geostatistics, which I leave to those knowledgeable

continued on page 12

How Many Samples Are Enough? continued from page 11

in the art. A recent example of a regional-scale soil study that took regionalized variability into consideration is the work of Grunsky et al. (2010). The impact for those collecting samples from a small area for an environmental study is that the true variance appropriate for the whole extent of a sampled media will likely be underestimated. Therefore, care has to be taken in extrapolating locally generated information to other generally similar areas to estimate just “how many samples should be enough” to satisfactorily sample a larger area. To use a geostatistical term, one should not extrapolate beyond the ‘support’ that the data in hand provide; which takes us back to where we started, proper sample design planning to meet the objectives of a survey.

References

- CANADIAN COUNCIL OF MINISTERS OF THE ENVIRONMENT (CCME). 1997. Canadian Soil Quality Guidelines for the Protection of Environmental and Human Health: Arsenic (Inorganic). <http://ceqg-rcqe.ccme.ca/download/en/257/>, visited September 30, 2010.
- GARRETT, R.G. 1979. Sampling considerations for regional geochemical surveys. In: Current Research Part A, Geological Survey of Canada Paper 79 1A, 197-205.
- GARRETT, R.G. 1983a. Sampling methodology. In: HOWARTH, R.J. (ed.) Handbook of Exploration Geochemistry, Vol. 2, Statistics and Data Analysis in Geochemical Prospecting, Elsevier, 83-110.
- GARRETT, R.G. 1983b. Opportunities for the 80's. Mathematical Geology, 15, 389-402.
- GARRETT, R.G., KANE, V.E. & ZEIGLER, R.K. 1980. The management and analysis of regional geochemical data. Journal of Geochemical Exploration, 13, 115-152.
- GRUNSKY, E.C., DREW, L.J. & SUTPHIN, D.M. 2010. Process recognition in multi-element soil and stream-sediment geochemical data. Applied Geochemistry, 24, 1602-1616.
- KOCH, G.S. & LINK, R.F. 1971. The coefficient of variation; a guide to the sampling of ore deposits. Economic Geology, 66, 293-301.
- REIMANN, C. & GARRETT, R.G. 2005. Geochemical background - Concept and reality. Science of the Total Environment, 350, 12-27.
- REIMANN, C., FILZMOSER, P. & GARRETT, R.G. 2005. Background and threshold: critical comparison of methods of determination. Science of the Total Environment, 346, 1-16.
- REIMANN, C., FILZMOSER, P., GARRETT, R. & DUTTER, R. 2008. Statistical Data Analysis Explained: Applied Environmental Statistics with R. John Wiley & Sons, Ltd., Chichester, England.
- ROCKE, D.M., DOWNS, G.W. & ROCKE, A.J. 1982. Are robust estimators really necessary? Technometrics, 24, 95-101.
- SMITH, D.B., WOODRUFF, L.G., O'LEARY, R.M., CANNON, W.F., GARRETT, R.G., KILBURN, J.E. & GOLDHABER, M.B. 2009. Pilot studies for the North American Soil Geochemical Landscapes Project - Site selection, sampling protocols, analytical methods, and quality control protocols. Applied Geochemistry, 24, 1357-1368.
- STANLEY, C.R. 2010. How many samples are enough? Explore, 148, 2-11.

Robert G. Garrett

Emeritus Scientist

Geological Survey of Canada

Natural Resources Canada

Ottawa, Ontario

CANADA, K1A 0E8

Email: garrett@NRCan.gc.ca

