Opinion: "Cotton is cotton, don't worry about sampling—just look at the data..."

Claudia Paoletti
GMO Unit, European Food Safety Authority (EFSA), Parma, Italy. E-mail: claudia.paoletti@efsa.europa.eu, http://www.efsa.europa.eu

At a recent collaborative session Claudi Paoletti expressed a long-standing frustration as to the reaction she has met with in the food, feed and commodities fields. An immediate invitation to air this frustration followed from the editor of *TOS forum*—et voila!

Some 15 years ago, when I asked for clarifications on how the plants I was assessing had been selected I was simply told: *"Cotton is cotton, don't worry about sampling—Just look at the data"*. One would wish this was a singular occasion, but I have spent many years in my professional life hearing the same thing over and over again: maize is maize, soybean is soybean... a kernel is a kernel... a seed is a seed: just look at the data (don't worry about all this sampling)". However, this clashed with everything I have ever been taught and studied: by definition, the process of sampling is always a source of error in itself when estimating population characteristics and when characterising heterogeneous lots. The Theory of Sampling (TOS) was developed specifically to define suitable strategies for obtaining reliable estimates from limited numbers of measurements, minimising the unavoidable sampling error. How was it possible that apparently *nobody* was worried about sampling when the focus was on obtaining those few cotton, soybean, maize plants/seeds from which I was presumed to make inferences of general relevance? A mystery!

Years later, in 2004, I decided to attend WCSB 2 in Perth, Australia. There I discovered that there were many scientists (prominently engineers, geologists and industry managers in the mining sectors) who were also worrying about sampling—who kindly introduced me to the Theory of Sampling (TOS), which started to shed some clarity on the many questions I had. Finally I was not alone anymore: indeed searching for diamonds in rocks, sediments and soils could not be so different from searching for defect kernels in a 60,000 tons shipment!

This boosted my motivation and when back in Europe I decided to carefully investigate standard sampling procedures for agricultural commodities. Several national and international organisations have developed and recommended approaches for kernel sampling (i.e. seeds and grains), including: the International Seed Testing Association (ISTA), the United States Department of Agriculture/Grain Inspection, Packers & Stockyards Administration (USDA/GIPSA), the *Comité Européen de Normalisation* (CEN), the WHO/FAO *Codex Alimentarius*, and the International Organization for Standardization (ISO).

The vast majority of the world's recommended sampling plans are based upon the fundamental *assumption* of a "random distribution" of the parameter of interest, so that the mean, the standard deviation of the mean and both the producer and consumer risks can be easily estimated according to the Binomial, the Poisson or the hypergeometric distribution. Nonetheless, assuming randomness without justification is very risky, if not completely wrong, as it has been demonstrated in specific cases. Experience shows that such "perfect disorder" in agricultural commodities is the exception, while partial order (i.e. strong irregular heterogeneity, spatially as well as compositionally) is rather the rule.

Industrial activities are operations narrowly defined and structured in time and space. This generates correlations that, among other consequences, promote segregation during transportation and handling of the material. In addition to the inherent heterogeneity in a population of natural units, e.g. a lot of particulate material (kernels), there is *always* also an amount of induced heterogeneity—for me it is therefore clear that assuming a random distribution is an irrational wish, not supported by empirical evidence. This convenient attitude simply encourages faulty solutions to sampling problems, overlooking the issue of heterogeneity. Experimental confirmation comes from several studies investigating the degree of heterogeneity for several traits in large seed lots. Extensive heterogeneity has also been reported for kernel lots produced with large-scale facilities, such as those for grass seed production in the

Figure 1. The perennial issue in science, technology and industry: grab sampling (because the lot *appears* to be homogenous). The worst approach to sampling ever!

mid-west US. A disturbing explanation was offered by some authors "*such seed lots are seldom if ever blended by state-of-the-art equipment, but are simply conditioned, bagged, and marketed*".

Sporadic attempts to adapt the mathematical properties of the Poisson distribution to events of non-random material distributions have been made in the past, but such approaches may well violate inherent assumptions (e.g. normal variance characteristics) required for the use of such tools and have not been pursued further.

Clearly, providing recommendations for sampling approaches suitable for agricultural commodities continues to be a challenge. On the one hand, a high likelihood of non-random distribution of contaminations in most market products must be expected. On the other, there is a lack of experimental data regarding the distribution of contaminants in the world's many different products. Yet, we know from TOS that the distribution of a contaminant in a bulk greatly affects the effectiveness of sampling procedures, indeed it may fatally dismiss any chance for representative sampling at all. It is clear, contrary to today's *status quo*, that an approach free of the constraint implicit in the assumption of random distribution is unavoidable.

A number of factors must be taken into account when defining sampling protocols. Among these, the definition of a maximum acceptable sampling error is of utmost importance. The degree of risk that both the consumer and the producer are prepared to accept in terms of getting a wrong result, will contribute to the definition of this maximum level threshold. Once this is fixed, the sampling protocol can be designed accordingly, so that the costs of a sampling survey can be minimised without compromising the reliability of the final analytical results beyond a certain level (the accepted risk).

Nevertheless, when sampling is executed to check for compliance with legislation requirements (i.e. regulatory sampling) it is of crucial importance to ensure a high degree of confidence that the survey is accurate (unbiased) and that the compound sampling error is as small as indeed possible, within specified economic and workload reasons. Specifically, if there is a legal threshold limit set for acceptance of the presence of a specific contaminant, all adopted sampling protocols must ensure that such threshold is respected with the specified degree of confidence. Of course, the lower this limit is, the greater the demands will be upon the sampling plans. Extensive results from both theoretical research as well as many experimental studies show unequivocally that heterogeneity rules with respect to contaminant distribution in bulk commodities. Together, these findings pose a serious limit to unconditional acceptance of the assumption of random distribution and to the use of a simplistic Binomial distribution to estimate producer and consumer risks.

So, where do we go from here? If providing reliable sampling recommendations is a priority for the scientific community, it is necessary to invest in research projects designed to collect data on real distributions in agricultural commodities, worldwide. This would allow proper calibration of the statistical models used to estimate the degree of expected lot heterogeneity, without relying on pure unfounded *speculations*.

Meanwhile, some precautions should be taken now. As raw materials often come from different suppliers and given that industrial operations are structured in space and time, we must expect that a vestige of the original chronological order will always present in the spatial heterogeneity of any lot. Under this assumption, a systematic sampling approach is to be preferred over a random one.

As far as the number of increments used to produce the bulk sample (the composite sample) is concerned, it is very difficult to make clear, general recommendations because the number of increments required to minimise the sampling error, according to some pre-defined expectation, will depend entirely on the effective heterogeneity of the lot under investigation. The severe lack of data on the expected distributions of real lots makes it impossible to establish objective *criteria* to broadly address this problem.

Unfortunately, representative sampling is often completely uncorrelated with sampling costs: a representative protocol will have a high cost in terms of both time and financial resources necessary to carry out the necessary sampling operation. Nevertheless, excuses to perform incorrect sampling can never be justified by time and money limitations. If the sampling process is not representative, there is no reason to carry out any sampling at all—the resulting analytical results will be fatally unreliable, because of the lack of acceptable evidence regarding the uncompromised field-to-aliquot pathway. These issues have been treated in full detail elsewhere.[1–3]

References

1. K.H. Esbensen, C. Paoletti and P. Minkkinen, "Representative sampling of large kernel lots—I. Theory of Sampling and variographic analysis", *Trends Anal. Chem. (TrAC)* **32,** 154–165 (2012). doi: http://dx.doi.org/10.1016/j.trac.2011.09.008
2. P. Minkkinen, K.H. Esbensen and C. Paoletti, "Representative sampling of large kernel lots—II. Application to soybean sampling for GMO control", *Trends Anal. Chem. (TrAC)* **32,** 166–178 (2012). doi: http://dx.doi.org/10.1016/j.trac.2011.12.001
3. K.H. Esbensen, C. Paoletti and P. Minkkinen, "Representative sampling of large kernel lots—III. General Considerations on sampling heterogeneous foods", *Trends Anal. Chem. (TrAC)* **32,** 179–184 (2012). doi: http://dx.doi.org/10.1016/j.trac.2011.12.002

Claudia Paoletti, PhD is deputy head of the GMO Unit at the European Food Safety Authority (EFSA), Parma, Italy.