Accred Qual Assur (2005) 10: 452–454
DOI 10.1007/s00769-005-0013-8

Hilko van der Voet

How to construct a confidence interval from only one measurement on a composite sample assuming log-normality and known variance for the increment samples

Received: 24 February 2005
Accepted: 10 July 2005
Published online: 22 November 2005

H. van der Voet (✉)
Biometris, Wageningen University and Research centre,
P.O. Box 100, NL-6700 AC
Wageningen, The Netherlands
e-mail: hilko.vandervoet@wur.nl

Keywords Sampling uncertainty · Composite sample · Increment sample · Lognormal distribution · Inspection

Introduction

You have to decide whether a lot of animal feed can be accepted on the basis of one analysis result. You know that the material is non-homogeneous, so that sampling uncertainty should be considered. There is some historical knowledge about the variability of the material: How to proceed?

For inspection purposes a lot is often sampled according to relevant regulations (e.g. EU directives 76/371 [1] or 98/53 [2]), by combining a number of increment samples into one composite (bulk or aggregate) sample, from which subsequently only one measurement of a characteristic of interest is obtained. An example is the official inspection of aflatoxin levels in lots of animal feed products.

For the purpose of comparing the measurement result with a regulatory limit value it is often necessary to specify the sampling uncertainty of the measurement result. In this note we will assume that sampling uncertainty overwhelms measurement uncertainty. Clearly, sampling uncertainty cannot be derived from one single value, and therefore we assume that knowledge is available about the heterogeneity of the characteristic from other, similar lots of material.

For many characteristics of interest, such as low-level chemical residue concentrations, the lognormal distribution is a sensible model to describe the distribution across the lot. The problem is that the measurement on the composite sample is no longer a direct observation from this lognormal distribution, but is the arithmetic mean of several observations from the same lognormal distribution.

For the situation where several composite samples are taken (from the same lot), it has been described how to use both mean and variance of the set of measurements to construct a confidence interval for the mean [3, 4]. However, for the much simpler situation of only one composite sample, but where the knowledge about variability of the lot is assumed to be known, no explicit expression of the confidence interval was found. In this note we derive this expression in a form suitable for practical use.

Method

Let us assume that a lot is one of a large number of lots for which the "same" heterogeneity can be assumed. We may characterise heterogeneity with relative measures, for example a relative standard deviation (RSD, defined as standard deviation divided by mean) or a geometric standard deviation (GSD, defined as antilog of the standard deviation of the logtransformed variable). These measures are typically quantified by measuring individual increment samples, and therefore describe the differences between portions of similar size as the increment samples. It is assumed in this note that a value of RSD or GSD is known. This may be a realistic value based upon a large amount of historical data or alternatively, it may be a subjective worst-case specification.

Suppose y is a single measurement result on a composite sample constructed from n increment samples. Then, with RSD or GSD for the increments known, an approximate 95% confidence interval for the true value μ_y of the

characteristic is

$$\frac{y}{k^2} \le \mu_y \le yk^2$$

with

$$k = \exp\left(\frac{RSD}{\sqrt{n}}\right) \quad \text{or} \quad k = \exp\left(\sqrt{\frac{GSD^{\ln(GSD)} - 1}{n}}\right) \tag{1}$$

A derivation of this result is given in the Appendix.

Example

In a study [5] increment samples were drawn from large shipments of copra meal pellets, copra cake and palm kernel cake to study the distribution of aflatoxin B_1. For 500 g increments in 16 lots, the within-lot heterogeneity was characterised on average by a relative standard deviation (RSD) of 22%. The maximum RSD was 40%. Therefore it seems reasonable to state that an RSD of 50% represents a conservative estimate.

Suppose that a composite sample is obtained by combining 20 increments of 500 g from a lot of 400 t palm kernel cake. The measured result is y=9.5 μg/kg aflatoxin B_1. Then, assuming that measurement uncertainty can be ignored, we obtain from Eq. (1) $k = \exp(0.50/\sqrt{20}) = 1.118$, and thus the following conservative 95% confidence interval

$$\frac{9.5}{1.118^2} \le \mu_y \le 9.5 \times 1.118^2$$

$$7.6 \le \mu_y \le 11.9.$$

Note that this interval is asymmetric around the point estimate y=9.5 μg/kg. Compared to a regulatory threshold of 5 μg/kg the interval shows a clear exceedance for aflatoxin B_1.

If the conclusion had been unclear (for example, against a regulatory threshold of 8 μg/kg), then we could expect a more narrow interval by the use of analysis results from the individual increment samples. We might expect a typical RSD of 22% and thus, via k=1.050, an interval $8.6 \le \mu_y \le 10.5$. If in such a case filed increment samples are still available, it would be worthwhile to analyse them separately.

Acknowledgement I thank Pieter Vereijken for his critical comments on the manuscript.

Appendix: derivation of Eq. (1)

If we denote the true value of an increment sample by x, then, ignoring measurement error, the composite sample value is:

$$y = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Assuming the same lognormal distribution for all x_i is the same, as assuming the same normal distribution for all $\ln(x_i)$. Let μ and σ^2 be the mean and variance of this normal distribution. The variance will be known and can be calculated from RSD or GSD as:

$$\sigma^2 = ln(1 + \text{RSD}^2) \quad \text{or} \quad \sigma^2 = \{ln\,(\text{GSD})\}^2 \tag{2}$$

The expected value of y, which is the same as the expected value of x_i, is:

$$\mu_y = \exp\left(\mu + \frac{1}{2}\sigma^2\right) \tag{3}$$

Equating the expected value μ_y to the observation y we obtain as an estimator of μ:

$$\hat{\mu} = \ln(y) - \frac{1}{2}\sigma^2 \tag{4}$$

The approximate variance of this estimator is (by the delta method):

$$\text{var}\,(\hat{\mu}) \approx \left(\frac{d\hat{\mu}}{dy}\right)^2 \text{var}\,(y) = \left(\frac{1}{y}\right)^2 \text{var}\,(y) \tag{5}$$

The standard formula for the variance of a lognormal observation is:

$$\text{var}\,(x) = e^{2\mu+\sigma^2}\left(e^{\sigma^2} - 1\right) \tag{6}$$

Combining Eqs. (5) and (6) gives:

$$\text{var}\,(\hat{\mu}) \approx \left(\frac{1}{y}\right)^2 \frac{1}{n}\text{var}\,(x) = \frac{1}{y^2 n} e^{2\mu+\sigma^2}(e^{\sigma^2} - 1)$$

Equating the expected value μ to the estimator $\hat{\mu}$ from Eq. (4) we obtain a simple expression for the estimated variance:

$$\hat{v}ar(\hat{\mu}) = \frac{1}{y^2 n} e^{2(\ln(y) - 1/2\sigma^2) + \sigma^2}(e^{\sigma^2} - 1) = \frac{e^{\sigma^2} - 1}{n}$$

An approximate $(1-2\alpha)$ confidence interval for μ is thus given by:

$$\hat{\mu} - z_\alpha\sqrt{\hat{v}ar\,(\hat{\mu})} \le \mu \le \hat{\mu} + z_\alpha\sqrt{\hat{v}ar\,(\hat{\mu})}$$

$$ln\,(y) - \frac{1}{2}\sigma^2 - z_\alpha\sqrt{\frac{e^{\sigma^2} - 1}{n}} \le \mu \le ln\,(y)$$

$$-\frac{1}{2}\sigma^2 + z_\alpha\sqrt{\frac{e^{\sigma^2} - 1}{n}}$$

where z_α is the (1−α)-point in the standard normal distribution (e.g. $z_{0.025}$=1.96). By adding $\frac{1}{2}\sigma^2$ and exponentiation (according to Eq. (3)) we obtain an approximate (1−2α) confidence interval for μ_y:

$$y\left\{\exp\left(\sqrt{\frac{e^{\sigma^2}-1}{n}}\right)\right\}^{-z_\alpha} \le \mu_y \le y\left\{\exp\left(\sqrt{\frac{e^{\sigma^2}-1}{n}}\right)\right\}^{z_\alpha}$$

Setting z_α to the approximate value 2, and replacing σ^2 with the appropriate expression for RSD or GSD (Eq. (2)), we obtain Eq. (1).

References

1. First Commission Directive 76/371/EEC of 1 March 1976 establishing Community methods of sampling for the official control of feedingstuffs
2. Commission Directive 98/53/EC of 16 July 1998 laying down the sampling methods and the methods of analysis for the official control of the levels for certain contaminants in foodstuffs
3. Zhou XH, Gao SJ (1997) Stat Med 16: 783–790
4. El-Baz A, Nayak TK (2004) Environ Ecol Stat 11: 283–294
5. Coker RD, Nagler MJ, Defize PR, Derksen GB, Buchholz H, Putzka HA, Hoogland HP, Roos AH, Boenke A (2000) J AOAC Int 83: 1252–1258