Parasitol Res (2004) 94: 361–366
DOI 10.1007/s00436-004-1225-2

ORIGINAL PAPER

David A. Morrison

Technical variability and required sample size of helminth egg isolation procedures: revisited

Received: 20 August 2004 / Accepted: 26 August 2004 / Published online: 30 September 2004

Abstract Mes [Vet Parasitol (2003) 115:311–320] recently reported a quantitative study of repeated measurements of nematode egg counts in faecal samples from dairy cattle, in order to compare the faecal egg counts resulting from two different laboratory techniques, the widely used McMaster method and a newer salt-sugar flotation (SSF) method. He concluded that the SSF technique requires much smaller sample sizes, and is also potentially simpler to carry out, making it the method of choice. Here I re-analyse these data to show that if the comparison is done on the most appropriate measurement scale (lognormal), and the large difference in multiplication factors is taken into account, then there is little to choose between the McMaster and SSF techniques as far as the required sample size is concerned. In particular, the treatment of the data as normally rather than lognormally distributed leads to incorrect statistical tests, power analyses and confidence intervals.

Introduction

Mes (2003) recently reported a quantitative study of data obtained from repeated measurements of nematode egg counts in faecal samples from dairy cattle. The objective was to compare the faecal egg counts (FEC), measured as eggs per gram (EPG), resulting from two different laboratory techniques, the widely used McMaster method and a newer salt-sugar flotation (SSF) method. This experiment allowed an assessment of the suitability of the two techniques in terms of their inherent technical variation (i.e. precision), and thus permitted recommendations to be made for the sample sizes required in order to estimate specified differences in EPG between experimental groups. It was concluded that the SSF technique requires much smaller sample sizes than does the McMaster technique when applied to the same faecal samples, with only half the size needed for samples with high numbers of eggs and one-fifth the size needed for samples with low numbers of eggs. Given that the SSF technique is also potentially simpler to carry out, this would make it the method of choice.

This is an important area of study, as true biological effects can only be detected in experiments that have been adequately designed with respect to sample size. However, I point out here that the experimental design used by Mes (2003) has a fundamental limitation, and that furthermore the data analyses employed are based on an inappropriate scale of measurement. The latter has caused the sample sizes to be mis-estimated for both techniques and thus the difference between the two techniques to be exaggerated. I present a re-analysis of the data on a more appropriate scale, and reach somewhat different conclusions about the relative merits of the two techniques.

Materials and methods

For the study of Mes (2003), a faecal sample was obtained from each of three calves that were expected to have different FECs: low (< 20 EPG), intermediate (50–200 EPG) and high (> 500 EPG). Each sample was divided into 20–24 aliquots for each method (McMaster and SSF), and analysed in the laboratory. The data obtained were then used in a retrospective power analysis (Sokal and Rohlf 1981) to estimate the number of technical replicates required by each of the two techniques for a range of expected EPG values. The sample sizes reported were 2–7 times greater for the McMaster technique compared to the SSF technique, and so Mes

D. A. Morrison (✉)
Department of Parasitology (SWEPAR),
National Veterinary Institute and Swedish University
of Agricultural Sciences, 751 89 Uppsala,
Sweden
E-mail: David.Morrison@vmm.slu.se
Tel.: +46-18-674161
Fax: +46-18-309162

(2003) concluded that the latter technique is to be preferred.

I re-analysed the published version of these data using a custom spreadsheet to log transformation the data (to fit them to a lognormal distribution) and to perform Welch and Student *t*-tests. Maximum likelihood analysis of the fit of the data to a lognormal distribution was performed using Regress+ version 2.3.1 (McLaughlin 1999). I re-calculated the power analyses of Mes (2003) using DSTPLAN version 4.2 (Brown et al. 2000), thus correcting some mis-calculations.

To compare means and variances with different multiplication factors, simulated random data were produced using SYSTAT version 9.01 (SPSS 1998), based on the mean and variance for the "Low FEC" samples listed by Mes (2003). Then the original continuous data were successively rounded to the nearest appropriate unit for the shown range of multiplication factors (i.e. 1, 2, 5, ..., 100), and the sample mean and variance were calculated separately for each factor.

Results and discussion

It is important to recognize that the most significant difference between the two techniques for estimating EPG is the effective size (usually volume or weight) of the faecal sample that is used in the laboratory analysis. The McMaster technique, as employed by Mes (2003), involved examining 0.02 g of faeces per aliquot, while the modified SSF technique involved examining 0.25 g. This means that the observed McMaster FEC for each aliquot must be multiplied by 50 in order to convert the aliquot counts to EPG, while the observed SSF FEC only needs to be multiplied by 4—this is called the multiplication factor or aliquot multiplier. The McMaster technique thus used a much coarser set of possible observed EPG values, as the values can only be 0 (zero eggs counted), 50 (one egg counted), 100 (two eggs), etc., while for the SSF technique they can be 0 (zero eggs counted), 4 (one egg counted), 8 (two eggs), etc.—this coarseness is usually referred to as the sensitivity or detection limit of the procedure. As a specific example, the 24 observations for the "Low FEC" McMaster sample shown by Mes (2003) are: 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 50, 50, 50. Clearly, these data represent a low detection limit and are not very sensitive to changes in egg abundance.

Unfortunately, it seems to be rarely appreciated that for any given faecal sample the multiplication factor used will affect the estimates of both the mean (i.e. the central location) and the variance (i.e. the spread) of the EPG values, irrespective of any other characteristics of the laboratory technique. This effect is shown using simulated data in Fig. 1. Clearly, as the coarseness of the scale (i.e. the size of the multiplication factor) increases, the estimated variance increases and the estimated mean decreases. For the mean, the decrease occurs irrespective of the sample size on which it is based, but is larger for smaller sample sizes (Fig. 1a). For the variance, the amount of increase depends directly on the sample size, reducing effectively to zero as the size increases (Fig. 1b). These two patterns have nothing whatever to do with any differences in the laboratory techniques used, but are theoretically inevitable results of the mathematical calculation of the averages and dispersions. That is, if the SSF technique examined only 0.02 g of faeces instead of 0.25 g then it too would have a larger variance and a reduced mean, all else being equal. The actual size of the difference in mean and variance will, however, depend on the FEC values, as the effect of the multiplication factor is small for large EPG values and increases as the EPG is lowered (Gill et al. 1986).

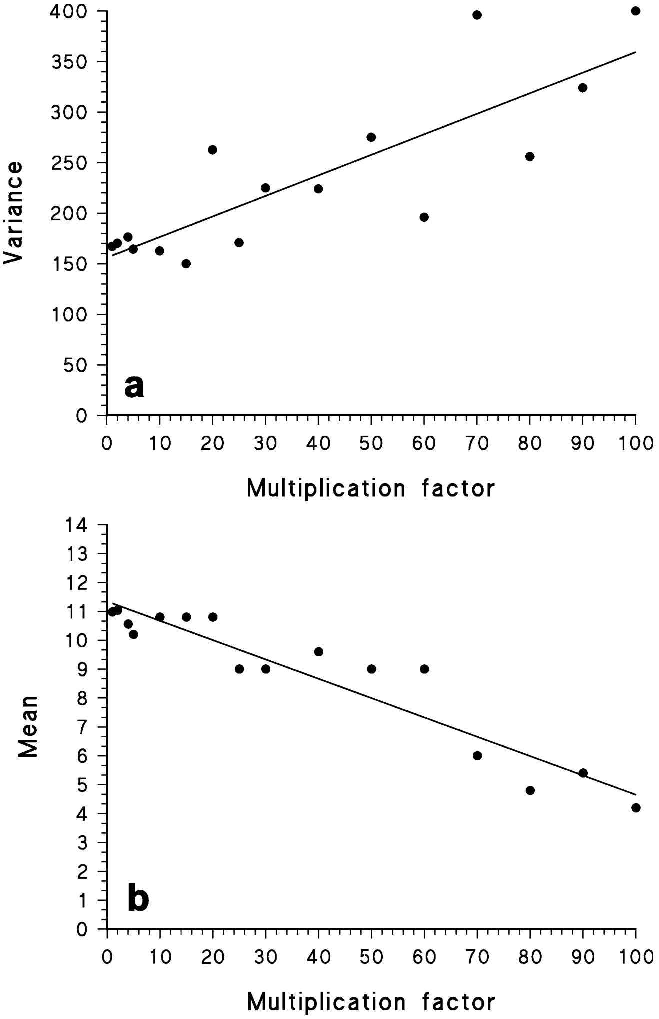

Fig. 1 Relationship between the multiplication factor needed to convert the original faecal egg count data of eggs per aliquot into eggs per gram and the **a** estimated sample mean and **b** estimated sample variance. For reference, the salt-sugar flotation technique has a multiplication factor of 4, while the McMaster technique has a multiplication factor of 50. The data points are for a single sample of 25 simulated observations from a lognormal frequency distribution, while the line is based on a sample of 10,000 such observations

Thus, direct comparisons of FEC data can only be validly made using the same multiplication factor—using different sensitivities means that the experiment is no longer a controlled one (i.e. the two treatments being compared are no longer the same in all ways except for the one feature being studied). For example, the comparison by Mes et al. (2001) of the SSF technique with the Wisconsin sugar flotation method is experimentally consistent, because a multiplication factor of 1 was used for both techniques, thus making the two sets of results directly comparable. On the other hand, the comparison by Mes (2003) of the McMaster and SSF techniques is confounded by the difference in multiplication factor. The consequence is that the differences in means and variances shown by Mes (2003) for the two techniques can be largely accounted for simply by the differences in the amount of faeces examined. We would expect a priori that the variances for the McMaster technique would be larger than the estimates for the SSF technique and that the means would be smaller. This is exactly what is shown by Mes (2003) for comparable samples, with the exception of the data for the "High FEC" means, which are confounded by a large (unexplained) difference in sample size.

Perhaps just as importantly for the conclusions reached by Mes (2003), the variation among samples all but disappears when the data are displayed and analysed on an appropriate measurement scale. The scale of measurement used to display and analyse quantitative observations is important because it determines the characteristics that the data are assumed to have, which in turn determine the outcome of the data analysis. If an inappropriate scale is used then incorrect conclusions can result (Sokal and Rohlf 1981). It is therefore necessary to decide, either in advance or after looking at the data when they are collected, what the most appropriate scale of measurement should be for the task at hand. It will frequently be found that biological data do not fit the usual linear scale of measurement, which is based on the idea that the data fit a normal frequency distribution.

FEC measured as EPG are analogous to a concentration—that is, they are a count per volume or weight. Concentrations are usually assumed to have a multiplicative scale of measurement, based on the idea that the data fit a lognormal frequency distribution (Limpert et al. 2001). For example, dilution series of chemicals are treated as multiplicative, as are dose-responses; and thus many biochemical measurements are most naturally measured on a logarithmic scale, such as pH, $p0_2$, pK and pI (in all cases the "p" indicates a negative logarithmic transformation of the original data). Therefore, we should expect a priori that EPG data will be most appropriately displayed and analysed using a logarithmic scale of measurement rather than a linear scale, because the influences on the intensity of parasite infection tend to be multiplicative rather than additive (Fulford 1994).

If this expectation is correct then we would predict that the estimated variance of each sample will be greater than the estimated mean, which is true for all of the data shown by Mes (2003). Furthermore, this expectation of a lognormal distribution can be confirmed by comparing the data on the linear scale with the same data after they have been ln transformed (Table 1). On the lognormal scale the variances are approximately equal, as confirmed by the variance-ratio tests, while on the normal scale they are very unequal in all cases. Thus, the empirical data of Mes (2003) meet the theoretical expectation very well.

As an aside, note that Mes (2003) suggests that the variability of faecal samples *between* animals (i.e. biological variation) is likely to follow a negative binomial frequency distribution rather than a normal distribution, because parasite infections are often clumped, so that relatively few hosts have most of the parasites. This may be true, as it has been commonly suggested for all sorts of species (Shaw et al. 1998), but it may not always be so, and the lognormal distribution has frequently been used instead (Gill et al. 1986). As an example comparison, the EPG data of Borgsteede et al. (2000) for 108 cows in the Netherlands fit a lognormal distribution (log likelihood, LL = –356.372) better than they do a negative binomial distribution (LL = –485.824) which in turn is a better fit than a normal distribution (LL = –669.663), as determined by maximum likelihood analysis. Nevertheless, the point being considered here is the variability between aliquots *within* a single faecal sample (i.e. technical variation), and I am arguing that this will usually be lognormally distributed rather than anything else.

The importance of a lognormal as opposed to a normal scale for the EPG data is that the differences in fineness of the scale, due to the different multiplication factors, will cause the estimates of the variance and mean to be exaggerated much more than they would be for data on a normal scale. This conclusion has several notable outcomes for the data analyses presented by Mes (2003).

First, the data analysed on the normal scale show a difference in estimated mean EPG for the "High FEC" samples only, while the data analysed on the lognormal scale show a difference in estimated mean EPG for the "Low FEC" samples only (Table 1). The latter result is much more in accord with expectations, as there is no a priori reason to expect statistically significant differences in the estimates of the two FEC techniques for samples with abundant parasite eggs, but there is good reason to expect a difference for samples with few eggs, since the coarseness of the McMaster scale will reduce the estimated mean (cf. Fig. 1a).

Second, the power analyses should all have been carried out using the data on the lognormal scale, since this is the most appropriate scale for the statistical analysis of any data collected. The re-calculated power analyses of Mes (2003) are shown in Table 2 along with the comparable analyses based on the lognormal scale. It is clear that the use of the normal scale causes the difference between the required sample sizes for the two techniques to be greatly exaggerated in all cases, by

Table 1 Descriptive and analytical statistics for the two faecal egg count (*FEC*) estimation techniques (McMaster and salt-sugar flotation, *SSF*) under different circumstances. For the normal measurement scale, the eggs per gram (*EPG*) data were retained on the original measurement scale, assuming that the frequency distribution of the observations was normal. For the lognormal measurement scale, the EPG data was ln transformed, assuming that the frequency distribution of the observations was lognormal. Variance test indicates the variance-ratio test for the equality of the variances for the paired samples from the McMaster and SSF techniques. Mean test at the normal measurement scale indicates the Welch *t*-test for the equality of the means for the paired samples from the McMaster and SSF techniques. For the lognormal scale the Student *t*-test for the equality of the means for the paired samples from the McMaster and SSF techniques was used

Technique	Normal measurement scale			Lognormal measurement scale		
	Variance	Variance test	Mean test	Variance	Variance test	Mean test
Low FEC (<20 EPG)						
McMaster	285.33	$F=4.88$	$t=2.05$	11.20	$F=1.40$	$t=2.69$
SSF	58.43	$P<0.001$	$P=0.052$	8.00	$P=0.213$	$P=0.010$
Intermediate FEC (50–200 EPG)						
McMaster	3,691.12	$F=6.85$	$t=0.24$	8.85	$F=1.29$	$t=1.28$
SSF	538.88	$P<0.001$	$P=0.809$	6.84	$P=0.274$	$P=0.207$
High FEC (>500 EPG)						
McMaster	103,421.05	$F=2.51$	$t=3.32$	6.20	$F=1.09$	$t=0.05$
SSF	41,153.28	$P=0.019$	$P=0.003$	5.70	$P=0.420$	$P=0.958$

Table 2 Calculated number of technical replicates required for the two FEC estimation techniques (McMaster and SSF) under different circumstances. Effect size indicates difference in average EPG between two populations of faecal samples, measured on the original measurement scale

Effect size (EPG)	Normal measurement scale		Lognormal measurement scale	
	McMaster	SSF	McMaster	SSF
Low FEC (<20 EPG)				
5	181	38[a]	69	50
10	46	11[a]	35	25
15	21	6[a]	25	19
Intermediate FEC (50–200 EPG)				
25	94	15	15	12
50	25	5	11	9
100	7	3	8	7
High FEC (>500 EPG)				
200	42	18	5	5
250	27	12	5	5
500	8	4	4	4

[a] Sample size wrongly estimated by Mes (2003), apparently based on a variance of 38.43 rather than 58.43

under-estimating the sample size for the SSF technique and over-estimating the sample size for the McMaster technique. Thus, the potential time-saving benefit of the SSF technique emphasized by Mes (2003) is to a large extent simply a product of the inappropriate choice of measurement scale rather than any inherent advantages of the laboratory technique itself. In particular, the difference between the two techniques does *not* extend to high levels of infection, as claimed by Mes (2003), but is restricted almost entirely to very low levels of infection. This presumably results from the fact that the multiplication factor has a bigger influence at low FEC levels than at higher levels (as noted above); and therefore this remaining difference is merely an artefact of the non-comparability of the two techniques as implemented by Mes (2003).

Third, it is worth pointing out that even though the McMaster technique, as employed by Mes (2003), requires 1–1.4 times as many technical replicates as does the modified SSF technique (Table 2), it actually involves examining much less faecal material than the SSF technique as a result of the smaller weight examined for each aliquot. Since the SSF technique examines 12.5 times as much faecal material per replicate, the amount of faeces examined for the complete laboratory analysis is actually 9–12.5 times as much for the SSF technique compared to the McMaster technique, for the same level of technical precision.

This raises the interesting question of how much more faecal material would need to be examined in the McMaster aliquots in order to be able to employ the same number of technical replicates as for the SSF technique. This is not easy to assess, as clearly it depends on the expected FEC in the sample, since for high expected values there is no difference between the results of the two techniques at all (and therefore no extra faecal material is needed). However, applying the simulation strategy used for Fig. 1 to the worst-case scenario shown in Table 2 (i.e. the top row of the "Low FEC" situation), and thus using 50 simulated samples from a lognormal frequency distribution, it can be estimated that a multiplication factor of 40 can reduce the variance to a similar value to that for a multiplication factor of 4, and a multiplication factor of 20 should do so most of the time.

The multiplication factor of any technique can be changed by varying the dilution of the faeces and/or the extract volume examined. So, the McMaster technique can easily be changed by adjusting the ratio of faeces to NaCl weight in the solution, and indeed this is what most of the published modifications to the McMaster technique are about. Furthermore, it is possible to obtain counting chambers of varying sizes (e.g. from Weber Scientific International, in Britain, and J.A. Whitlock & Co., in Australia) in order to adjust the volume of mixture actually examined for eggs, and therefore the multiplication factor. Thus, the precision of the two FEC techniques can easily be made comparable. In fact, many authors actually use a sensitivity of 25 rather than 50 (e.g. Borgsteede et al. 2000), and clearly this can be recommended based on the above calculations.

It thus seems that, if the comparison is done on the most appropriate measurement scale (lognormal), and the large difference in multiplication factors is taken into account, then there is little to choose between the McMaster and SSF techniques as far as the required sample size is concerned, at least for the cattle faecal data presented by Mes (2003). This conclusion is quite different from that of Mes (2003).

Furthermore, please realize that it would be unwise to use different multiplication factors for different samples in a single experiment (e.g. based on the expected FEC of the sample), because of the known effect on the estimated mean (cf. Fig. 1a). The data will not be directly comparable if different samples have different sensitivities. Thus, a single multiplication factor should be chosen, and then used for all of the data that are to be compared in any one experiment. Furthermore, it is quite likely that the high variability of FECs reported in the parasitology literature is due to the different sensitivities used by different experimenters, as this will be confounded with the different experimental conditions studied.

Importantly, note that use of the lognormal scale means that the best estimate of “average” EPG for a single faecal sample is usually the geometric mean of the repeated measurements rather than the arithmetic mean (Gill et al. 1986; Fulford 1994; Smothers et al. 1999). This is in contrast to many other frequency distributions (including both the normal and the negative binomial), where the arithmetic mean will be the same as the expected value (i.e. in the formal mathematical sense of moments). One important reason for this, which is often overlooked, is that confidence intervals for the lognormal distribution will not be symmetrical about the expected value (because the frequency distribution itself is not symmetrical, being bounded by zero on one side and infinity on the other). This asymmetry is not taken into account by the arithmetic mean but it is for the geometric mean, as can be seen in Table 3 for the data of Mes (2003). In this case, the 95% confidence interval for the EPG sample for the McMaster technique at “Low FEC” overlaps zero, indicating that we can be 95% confident that the true EPG average for this sample could be a negative number. Since this claim is clearly nonsensical from a biological perspective, the arithmetic mean calculations are of no practical value under these circumstances, and the geometric mean and confidence interval are to be preferred.

Unfortunately, if there are any zero values in the sample then by definition the geometric mean will always be zero. This is usually dealt with by adding 1 to each observation, performing the calculations, and then subtracting 1 from the result (Gill et al. 1986; Smothers et al. 1999). This is a perfectly adequate procedure if there are only a few 0 observations in the sample, but if there are many such observations then the value added will have an appreciable effect on the resulting estimate of the geometric mean. For example, if we use a $\ln(x+0.1)$ transformation instead of $\ln(x+1)$ then the geometric mean and confidence interval for the EPG sample for the McMaster technique at “Low FEC” changes from that shown in Table 3 to: mean = 0.22, 95%CI = 0.09–0.53. Therefore, you need to be cautious when comparing samples with zeroes to those without. Alternatively, Cox et al. (2000) recommend a different (and possibly better) procedure for dealing with zero counts. These points emphasize that we always need to consider carefully the nature of the data before interpreting the output from any data analyses.

In conclusion, I have made the following points:

1. The experimental design used by Mes (2003) does not easily allow a direct comparison of the McMaster and SSF techniques because of the use of different mul-

Table 3 Descriptive statistics for the arithmetic and geometric means for the two FEC estimation techniques (McMaster and SSF) under different circumstances

Technique	Arithmetic measurement scale[a]			Geometric measurement scale[b]		
	Mean	Lower CI[c]	Upper CI[d]	Mean	Lower CI[c]	Upper CI[d]
Low FEC (< 20 EPG)						
McMaster	6.25	–0.88	13.38	1.63[e]	0.93	2.86
SSF	14.00	10.77	17.23	12.29[e]	9.90	15.25
Intermediate FEC (50–200 EPG)						
McMaster	72.92	47.27	98.57	56.02[e]	41.23	76.12
SSF	76.17	66.13	86.21	72.86	64.05	82.88
High FEC (> 500 EPG)						
McMaster	1450.00	1299.49	1600.51	1415.60	1277.62	1568.49
SSF	1174.33	1088.67	1259.99	1157.19	1076.36	1244.09

[a] Arithmetic mean (i.e. based on untransformed data)

[b] Geometric mean (i.e. based on ln-transformed data)

[c] Lower boundary of the 95% confidence interval

[d] Upper boundary of the 95% confidence interval

[e] These data contain some values of 0, and the so the geometric mean is not unique. In this example, the data are effectively calculated using an $\ln(x+1)$ transformation rather than $\ln(x)$

tiplication factors, which affects the estimation of the means and variances.

2. The treatment of the data as normally distributed rather than lognormally distributed exaggerates the apparent differences between the results of the two techniques resulting from the above point.
3. The treatment of the data as normally distributed rather than lognormally distributed leads to incorrect statistical tests (showing a difference between the techniques at "High FEC" rather than "Low FEC"), power analyses (which over-estimate the sample size needed for the McMaster technique and under-estimate that needed for the SSF technique) and confidence intervals (which inappropriately overlap zero).
4. There seems to be little to choose between the two techniques when the above points are considered together, at least for the dataset under discussion.

As an addendum, it is worth noting that, in places, Mes (2003) inappropriately extrapolates the analysis of technical variation (i.e. variation among technical replicates from a single faecal sample) to biological variation among animals (i.e. variation among faecal samples taken from different individuals). This is unjustified, both because no data are presented on the nature of the latter variation and because it is often assumed that the data are likely to be most appropriately measured on a different scale (i.e. negative binomial rather than lognormal). The distinction between technical and biological variation is an important one (the first affects precision and the second accuracy), and both sources of variation need to be effectively dealt with in any worthwhile experiment. A quantitative analysis of between-animal variation would therefore be a useful adjunct to this work, allowing proper estimates of the number of animals needed for replication of experimental groups. Gill et al. (1986) have made an initial theoretical stab at this.

Acknowledgements Thanks to Ted Mes for helpful discussion and comment concerning the work presented here, and to Arvid Uggla for commenting on the manuscript.

References

Borgsteede FHM, Tibben J, Cornelissen JBWJ, Agneessens J. Gaasenbeek CPH (2000) Nematode parasites of adult dairy cattle in the Netherlands. Vet Parasitol 89:287–296

Brown BW, Brauner C, Chan A, Gutierrez D, Herson J, Lovato J, Polsley J, Russell K, Venier J (2000) DSTPLAN: Calculations for sample sizes and related problems. M.D. Anderson Cancer Center, Department of Biomathematics, University of Texas, Houston, Tex., http://odin.mdacc.tmc.edu/anonftp

Cox JL, Heyse JF, Tukey JW (2000) Efficacy estimates from parasite count data that include zero counts. Exp Parasitol 96:1–8

Fulford AJC (1994) Dispersion and bias: can we trust geometric means? Parasitol Today 10:446–448

Gill JL, Ericsson GF, Helland IS (1986) Precision of assessing anthelmintic efficacy. Biometrics 42:981–987

Limpert E, Stahel WA, Abbt M (2001) Log-normal distributions across the sciences: keys and clues. BioScience 51:341–352

McLaughlin MP (1999) Regress+: a tool for mathematical modeling. McLean, Virg. http://www.causascientia.org/software/Regress_plus.html

Mes THM (2003) Technical variability and required sample size of helminth egg isolation procedures. Vet Parasitol 115:311–320

Mes THM Ploeger HW, Terlou M, Kooyman FNJ, Van der Ploeg MPJ, Eysker M (2001) A novel method for the isolation of gastro-intestinal nematode eggs that allows automated analysis of digital images of egg preparations and high throughput screening. Parasitology 123:309–314

Shaw DJ, Grenfell BT, Dobson AP (1998) Patterns of macroparasite aggregation in wildlife host populations. Parasitology 117:597–610

Smothers CD, Sun F, Dayton AD (1999) Comparison of arithmetic and geometric means as measures of a central tendency in cattle nematode populations. Vet Parasitol 81:211–224

Sokal RR, Rohlf FJ (1981) Biometry, 2nd edn. Freeman, New York

SPSS (1998) SYSTAT 9 for Windows. SPSS, Chicago. http://www.systat.com/