Institute of Parasitology, Biology Centre CAS
Folia Parasitologica 2017, 64: 025
doi: 10.14411/fp.2017.025

Folia Parasitologica
http://folia.paru.cas.cz

Research Article OPEN ACCESS

Do small samples underestimate mean abundance? It depends on what type of bias we consider

Jenő Reiczigel[1] and **Lajos Rózsa**[2,3]

[1] Department of Biomathematics and Informatics, University of Veterinary Medicine, Budapest, Hungary;

[2] Hungarian Academy of Sciences, MTA-ELTE-MTM Ecology Research Group, Budapest, Hungary;

[3] Hungarian Academy of Sciences, MTA Centre for Ecological Research, Evolutionary Systems Research Group, Tihany, Hungary

Abstract: Former authors claimed that, due to parasites' aggregated distribution, small samples underestimate the true population mean abundance. Here we show that this claim is false or true, depending on what is meant by 'underestimate' or, mathematically speaking, how we define 'bias'. The 'how often' and 'on average' views lead to different conclusions because sample mean abundance itself exhibits an aggregated distribution: most often it falls slightly below the true population mean, while sometimes greatly exceeds it. Since the several small negative deviations are compensated by a few greater positive ones, the average of sample means approximates the true population mean.

Keywords: sampling bias, sample size, quantitative parasitology, aggregated distribution

In case of natural infections, parasites always exhibit an aggregated (right-skewed) distribution across host individuals: most hosts harbour a few if any parasites, while a few hosts harbour many parasite individuals (Crofton 1971). This distribution pattern influences several aspects of the ecology and evolution of host-parasite interactions (Krasnov 2008, Poulin 2011, Schmid-Hempel 2011, Clayton et al. 2016), and it may also cause a plethora of methodological problems in sampling design and statistical analysis (Rozsa et al. 2000).

In their pioneering study, Gregory and Woolhouse (1993) analysed how small sample size may bias the sample estimates of the population mean abundance (as defined by Bush et al. 1997) and other statistical measures of parasite burdens. Using computer simulations, they found that the smaller the sample size, the more often sample mean abundance underestimates the true population mean abundance. Intuitively, smaller samples are more likely to contain only individuals from the non-infected and slightly infected majority of the whole population, while the few heavily-infected individuals tend to be totally absent from small samples. They interpreted these results by concluding that "as sample size decreases values of sample mean parasite burden (…) are (…) systematically underestimated". Several subsequent authors (see e.g. Poulin 1996, Cunha-Barros et al. 2003, Marques and Cabral 2007, Mladieno et al. 2012) have reached similar conclusions. The purpose of our present account here is to explain that this is an over-interpretation of results. The bias of an estimate can be defined in several ways, of which two widely used definitions will be considered below. When speaking about underestimation or bias, one must clarify in which sense the statement is meant.

In mathematical statistics, an estimator is called 'unbiased' if the estimates are on average equal to the population parameter of interest. The term 'on average' is meant here in an abstract sense, that is taking (virtually) the average of all possible samples (which is typically an infinite set). The mathematical notion for this theoretical average is the 'expectation' or 'expected value'. Practically, unbiasedness means that taking a large number of random samples, the average of the sample estimates approximates the population parameter. However, if an estimate has a right-skewed sampling distribution, its unbiasedness implies that more than 50% of the estimates are smaller than the population parameter. Typically, there are several small negative deviations and a few large positive ones, so the median of the estimates is located below the population parameter. If it is required that 50% of the estimates lie left and 50% right of the population parameter, other estimators have to be used, which are not unbiased in the above sense.

As there have been many cases in which just this was required, also this property got its own name: this is the so-called 'median unbiasedness'. The difference between unbiasedness and median unbiasedness lies in the fact that the average is sensitive to the magnitude of deviations while

Address for correspondence: L. Rózsa, MTA-ELTE-MTM Ecology Research Group, H-1117 Budapest, Pazmany s. 1/C, Hungary.
Phone: +3612101075; Fax: +3613171669; E-mail: lajos.rozsa@gmail.com

the median ignores the magnitude, just counts the positive and negative deviations. A sufficient condition for an estimator to be both unbiased and median unbiased at the same time is that its sampling distribution is symmetrical.

Switching back to Gregory and Woolhouse (1993), they were right to point out that the majority of small samples underestimate the true population mean abundance, and only a minority of them will overestimate it. Unfortunately, they reported this result by claiming that population mean abundance is 'systematically underestimated' without specifying what does this exactly mean. Taking the sample from a host population exhibiting an aggregated (right-skewed) distribution of parasites, the sample mean abundance itself shows an aggregated distribution. This means that most sample mean abundances slightly underestimate the true population mean abundance, while a few of them greatly overestimate it. Thus the theoretical mean of the sample means equals the true population mean abundance: the sample mean is an unbiased estimate of the population mean. (Note that unbiasedness of the sample mean can also be proven theoretically, irrespective of the distribution of data.) Contrarily, the median of the sample mean abundances indeed underestimates the true population parameter, particularly in case of small sample sizes.

If the sample size tends to infinity, the distribution of sample means tends to the normal distribution. This phenomenon is expressed mathematically by the 'central limit theorem' (Rice 2007). Since the median of the normal distribution is equal to its mean, increasing the sample size reduces the difference between the mean and median of the sample mean abundance, and in limit the difference vanishes.

MATERIALS AND METHODS

Please note that below the terms 'mean', 'median', 'sample size' and 'distribution' are used on three different levels: (i) simulated population of hosts, (ii) samples of hosts derived from the simulated population, and (iii) means of these samples. To avoid potential confusions, we do our best to separate these measures as clearly as possible.

To illustrate the different effects of small sample sizes on the mean versus the median of sample means, we made a computer simulation using R 3.0.2 (R Core Team 2013). Using R functions written by ourselves, we created a virtual population of hosts (N = 100,000), harbouring parasites exhibiting a negative binomial distribution (a widely used mathematical model for aggregated parasite distributions) with mean abundance = 10 and exponent k = 0.05. This means that our virtual host-parasite system exhibited a highly aggregated distribution of parasites across host individuals. In the whole host population, 88% of individuals carried infections lighter than the population mean abundance. Then we took 10,000 random samples (sampling with replacement) of different sample sizes from the above host population.

RESULTS

We show results for arbitrarily chosen sample sizes (10, 30, 100 and 300) that roughly cover the range of sample sizes typically used in most practical studies. For sample sizes of 10, 30, 100 and 300, the sample mean abundance was

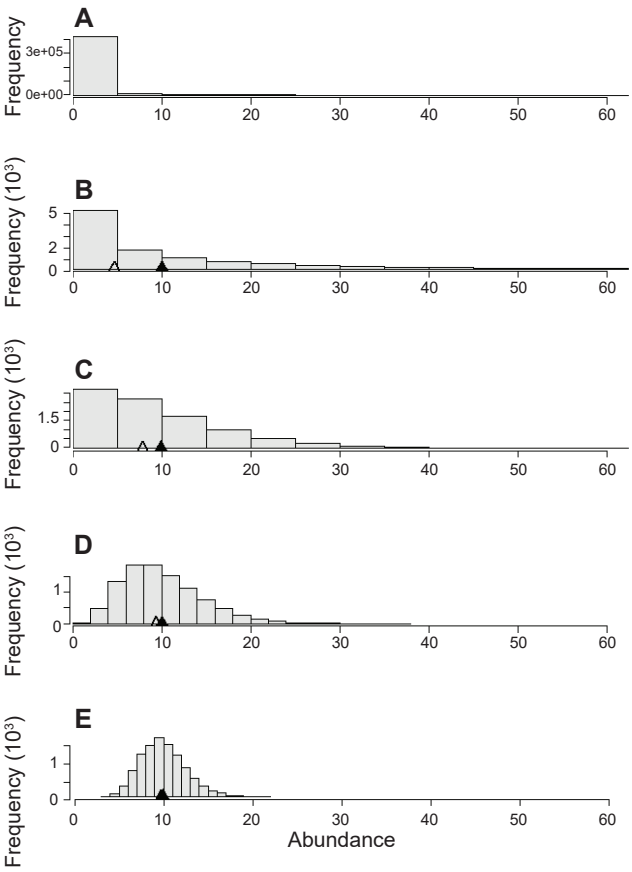

Fig. 1. A – a virtual population of hosts harbouring parasites exhibiting a highly aggregated negative binomial distribution (N = 100,000, mean abundance = 10, exponent k = 0.05) used as a model system; **B**–**E** – the distribution of sample mean abundances for sample sizes of 10, 30, 100 and 300 (each based on 10,000 random samples). For small samples the median of sample means (open triangle) greatly underestimates the true population abundance, whereas their mean (closed triangle) does not. With increasing sample size, the distribution becomes more and more symmetric, and the difference between the median and the mean of sample estimates vanishes.

below the population mean in 68%, 61%, 56% and 53% of the samples, respectively. This indicates that sample mean abundance values themselves were aggregated, although less aggregated than the distribution of abundance in the whole population. Further, this bias was more emphasised at small sample sizes, and the distribution of sample mean abundance values tended toward a normal distribution (but did not reach it) when we increased sample size (Fig. 1). Moreover, in Fig. 2 we also illustrate how the mean, median, 5% and 95% quantiles of the sample mean abundances depended on sample size. Please note that the sampling bias phenomena discussed above are caused exclusively by the aggregated nature of the distribution we modelled. Therefore, arguably, other aggregated distributions would yield qualitatively similar results.

Finally, due to the asymmetry of distribution, symmetrical CIs based on normal distribution do not work well for aggregated parasite distributions. For the virtual population we used in the present study, the actual coverage at a nominal level of 95% was 84%, 88% and 90% for sample sizes

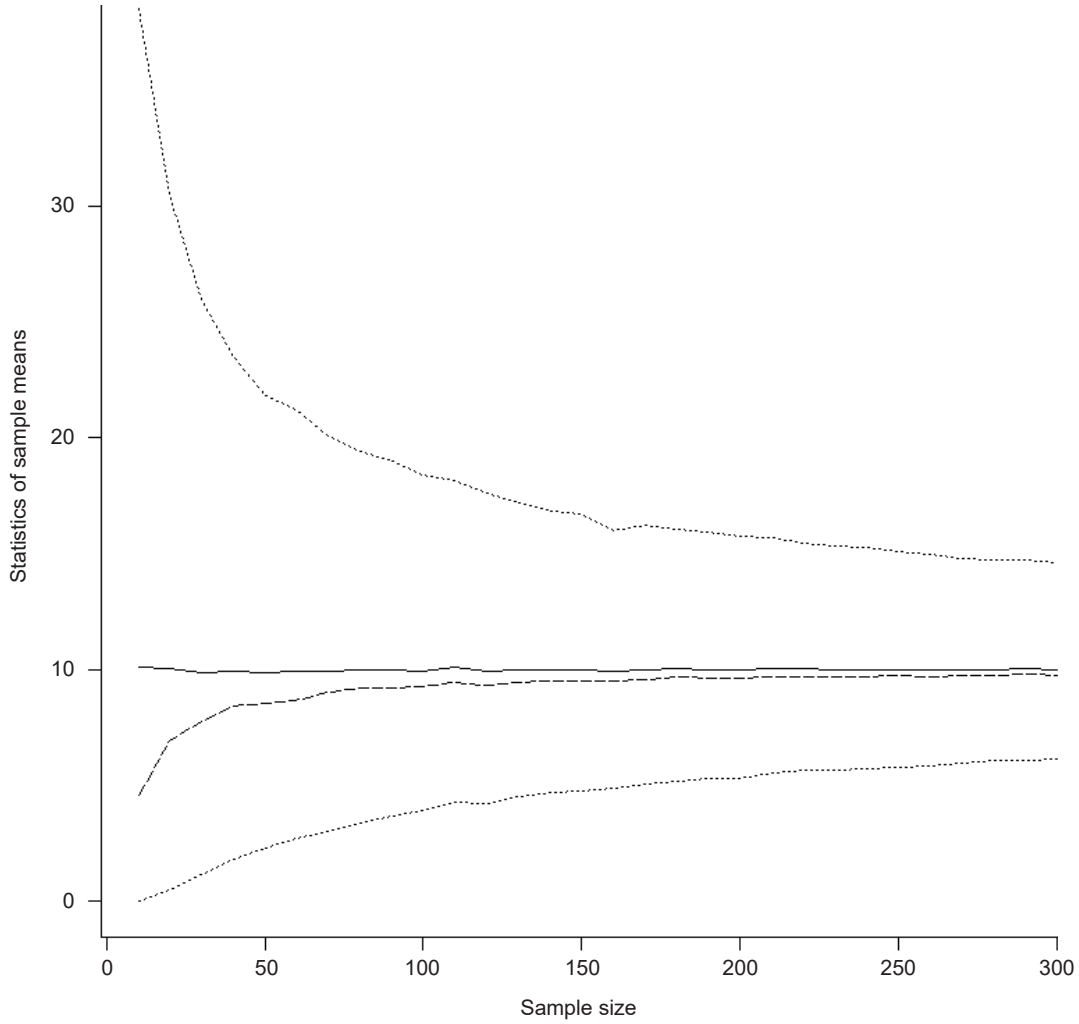

Fig. 2. Ten thousand random samples were taken (with replacement) from the virtual population shown in Fig. 1, and the mean (solid), median (dashed), and the 5% and 95% quantiles (dotted) of sample mean abundances are depicted as functions of sample size. The mean of sample means is very close to 10, the true population mean abundance, even for small samples. Contrarily, the median of sample mean abundances clearly underestimates it. The quantiles show the asymmetry of the distribution, which decreases with increasing sample size.

n = 100, 200 and 300, respectively. A usual solution to obtain better CIs is applying the bootstrap that is supposed to result in coverage probabilities closer to the nominal. Best coverage can be expected from the bias corrected (BCa) CI proposed by Efron (1987), which resulted in coverage probabilities of 90%, 92% and 93% for n = 100, 200 and 300, respectively, when applied to our data. This means that for reliable CI construction one needs large samples if the distribution is extremely aggregated.

DISCUSSION

To our best knowledge, Poulin (1996, 2011) published the only simulation study that contradicted our present conclusions. He found that the mean of sample mean abundances consistently underestimated the true population mean abundance at small sample sizes. Through private communications, however, the author informed us that those simulations were carried out using a software that could not be retrieved later on. Presumably, its results were caused by unknown artefacts, e.g. due to the small number (25–25) of samples in each sample size categories.

In conclusion, Gregory and Woolhouse (1993) – and several further authors following them – were correctly describing the 'median bias' phenomenon, more sample means falling below the population mean then above it. While this may be important to consider when estimating the population mean abundance from small samples, we caution against over-interpreting this phenomenon. Whether or not small samples tend to underestimate the population mean abundance depends on which type of bias we are talking about.

Acknowledgements. Thanks to Robert Poulin for private correspondence on this subject. Our work was supported by the grants from the National Scientific Research Fund of Hungary (OTKA/NKFI grant no. 108571) and GINOP-2.3.2-15-2016-00057.

REFERENCES

Bush A.O., Lafferty K.D., Lotz J.M., Shostak A.W. 1997: Parasitology meets ecology on its own terms: Margolis et al. revisited. J. Parasitol. 83: 575–583.

Clayton D.H., Bush S.E., Johnson K.P. 2016: Coevolution of Life on Hosts: Integrating Ecology and History. University of Chicago Press, Chicago, 294 pp.

Crofton H.D. 1971: A quantitative approach to parasitism. Parasitology 62: 179–193.

Cunha-Barros M., Van Sluys M., Vrcibradic D., Galdino C.A.B., Hatano F.H., Rocha C.F.D. 2003: Patterns of infestation by chigger mites in four diurnal lizard species from a restinga habitat (Jurubatiba) of Southeastern Brazil. Braz. J. Biol. 63: 393–399.

Efron B 1987. Better bootstrap confidence intervals. J. Am. Stat. Assoc. 82: 171–185.

Gregory R.D., Woolhouse M.E.J. 1993: Quantification of parasite aggregation: a simulation study. Acta Trop. 54: 131–139.

Krasnov B.R. 2008: Functional and Evolutionary Ecology of Fleas: a Model for Ecological Parasitology. Cambridge University Press, Cambridge, 593 pp.

Marques J.F., Cabral H.N. 2007: Effects of sample size on fish parasite prevalence, mean abundance and mean intensity estimates. J. Appl. Ichthyol. 23: 158–162.

Mladineo I., Šimat V., Miletić J., Beck R., Poljak V. 2012. Molecular identification and population dynamic of *Anisakis pegreffii* (Nematoda: Anisakidae Dujardin, 1845) isolated from the European anchovy (*Engraulis encrasicolus* L.) in the Adriatic Sea. Int. J. Food Microbiol. 157: 224–229.

Poulin R. 1996: Measuring parasite aggregation: defending the index of discrepancy. Int. J. Parasitol. 26: 227–229.

Poulin R. 2011: Evolutionary Ecology of Parasites. Princeton University Press, Princeton, 332 pp.

R Core Team 2013: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. World Wide Web electronic publication, www.R-project.org, 04/2017.

Rice J.A. 2007: Mathematical Statistics and Data Analysis. Thomson Higher Education, Belmont, 603 pp.

Rozsa L., Reiczigel J., Majoros G. 2000: Quantifying parasites in samples of hosts. J. Parasitol. 86: 228–232.

Schmid-Hempel P. 2011: Evolutionary Parasitology: the Integrated Study of Infections, Immunology, Ecology, and Genetics. Oxford University Press, Oxford, 516 pp.

Received 28 April 2017 Accepted 29 June 2017 Published online 26 July 2017

Cite this article as: Reiczigel J., Rózsa L. 2017: Do small samples underestimate mean abundance? It depends on what type of bias we consider. Folia Parasitol. 64: 025.