

Does the Pareto Principle describe parasite counts in humans? The case of hookworm in pregnant women

Yenifer Orobio^{1 2¶}, Neal Alexander^{1 2*}

¹ Centro Internacional de Entrenamiento e Investigaciones Médicas, CIDEIM, Cali, Colombia

² Universidad Icesi, Cali, Colombia

¶ ORCID 0000-0002-2103-4969

*Corresponding author

E-mail: nalexander@cideim.org.co (NA), ORCID 0000-0002-6707-7876

Abstract

The Pareto Principle (PP) — that 80% of an attribute are found in 20% of individuals — is one way to characterize heterogeneity in infectious diseases. An alternative is the dispersion parameter (k) of the negative binomial distribution (NBD). The NBD has two parameters, while the PP is a single property which may hold for any distribution family. The objectives of the current work are: a) to obtain a relation between the PP and NBD, i.e. to specify which combinations of NBD parameters comply with the PP; b) for hookworm, a soil-transmitted intestinal parasite of humans, to identify whether the PP or the NBD is a more parsimonious description of heterogeneity of infection load. For objective a), an empirical relation is found in the form of a saw-toothed line on a plot of k against the mean, reaching an asymptote of approximately 0.24. For objective b), we estimate k and the mean from nine studies from a systematic review of hookworm in pregnant women. Seven studies had higher heterogeneity than the PP, ranging from 83:20 to 100:20: we call these super-Pareto. One study was sub-Pareto (74:20), and one was Pareto (80:20). This suggests that at least two parameters, as supplied by the NBD, are necessary to describe the heterogeneity of hookworm. The probability of reaching a target reduction in prevalence is less when there is greater aggregation, which suggests that estimating aggregation via a subsample could be worthwhile, in order to set a target coverage threshold before starting mass drug administration.

Key words: helminths, heterogeneity, negative binomial distribution, Pareto principle, public health, control strategies

Introduction

Infectious disease processes are typically heterogeneous, as reflected in phenomena such as ‘superspreaders’, i.e. a small proportion of hosts being responsible for a large proportion of transmission (Jankowski et al. 2013; Lloyd-Smith et al. 2005; Stein 2011). One way to characterize such heterogeneity is the Pareto principle, also known as the 80/20 rule, which states that 80% of a certain characteristic is harbored by 20% of the population (Anderson and May 1991; Hardy 2010; Holland 2013). This has been used to describe the distribution of transmission and spread of infection among hosts of a variety of diseases (Cooper et al. 2019; Stein 2011; Woolhouse et al. 1997). In the case of count data, another way to characterize heterogeneity is via the dispersion parameter (k) of the negative binomial distribution (NBD), which is a generalization of the Poisson (Lloyd-Smith 2007). The NBD has often been found to well describe host-level patterns of helminths and other parasites (Anderson and May 1991). The NBD has two parameters (k and the mean, μ), while the PP is a single numerical property which may hold for any distribution family.

In the current paper we compare the ability of the PP and the NBD to characterize heterogeneity in infection intensity of hookworm, specifically hookworm in pregnancy. Hookworm is a soil-transmitted intestinal helminthiasis, which the World Health Organization (WHO) considers a major cause of morbidity, closely linked to poverty and inadequate personal hygiene, consumption of raw foods, fecal contamination of the environment and lack of sanitary services and drinking water (Martínez de la Ossa et al. 2010; Pawlowski et al. 1992; Valerio et al. 2002). The human hookworm pathogens are *Necator americanus* and *Ancylostoma duodenale*. The resulting morbidity is associated with infection intensity (Organización Panamericana de la

Salud 2012), so epidemiological descriptions need to reflect the distribution of numbers of parasite between hosts, as well as prevalence. Intense infections cause chronic blood loss that can lead to physical, nutritional and cognitive deterioration, specifically growth retardation and anemia in children and pregnant women (Organización Mundial de la Salud 2005; Organización Panamericana de la Salud 2012; Roca et al. 2003).

The objectives of the current work are: a) to obtain a relation between the PP and NBD, i.e. to specify which combinations of NBD parameters comply with the PP; b) to determine, in the case of hookworm in pregnancy, whether the PP or the NBD is a more parsimonious description of heterogeneity.

Materials and Methods

Comparison of the negative binomial distribution and the Pareto principle

Some combinations of NBD parameters give a lower degree of heterogeneity, e.g. 70:20 rather than 80:20, and we call this *sub-Pareto*. Other combinations give a higher degree, e.g. 90:20, which we call *super-Pareto*, while we will call a distribution in accordance with the Pareto principle *equi-Pareto*. The strategy of the current paper is to fit the negative binomial to observed data, read off whether this fitted distribution is *sub-*, *super-* or *equi-Pareto*, and so determine whether or not the Pareto is a parsimonious description of the heterogeneity of the data.

Negative binomial distributions which satisfy the Pareto Principle

We use the following procedure to identify those combinations of the negative binomial dispersion (k) and mean (μ) parameters which meet the PP. Values of μ were considered between 1 and 5000: more specifically, a geometric series of 1000 values between these limits, i.e. the log-values were uniform between $\log(1)$ and $\log(5000)$. Then, for each value of μ , the value of k was identified which satisfied the PP. To do so, the ‘nlminb’ function of R optimized over the proportion of the total count found in the highest 20 percent of the values. The cumulative distribution of the negative binomial was obtained as the cumulative sum of its probability function (R Core Team 2019; Venables and Ripley 2002). Hence we define an empirical, non-algebraic, relation between those values of μ and k which satisfy the PP constraint.

Case study: hookworm in pregnancy

Hookworm in humans is caused by two species of nematode parasites: *Necator americanus* and *Ancylostoma duodenale*, whose larvae infect people via contact with soil (Hotez et al. 2004). Disease results largely from blood loss caused by the adult parasites in the intestine. The morbidity largely results from anemia, and is associated with the intensity of infection. Almost 500 million people in developing tropical countries are infected (Loukas et al. 2016), including a quarter to a third of pregnant women in sub-Saharan Africa. The latter estimates are from a systematic review of hookworm-related anemia among pregnant women (Brooker et al. 2008), whose data we use in the current paper. The variables extracted from the systematic review were: sample size of the study population, proportion of people with infection, initial prevalence of infection, infection intensity category.

The original studies report hookworm eggs per gram (epg), which are obtained by applying a conversion factor to the numbers of eggs counted under the microscope. For example, one Kato Katz slide usually contains approximately 1/24 g of feces, so, if basing the eggs per gram on a single slide, the egg count would be multiplied by 24 (Organización Mundial de la Salud 1994; World Health Organization 1992). To apply the NB distribution, we need to infer the original egg counts which take values 0, 1, 2, etc., as opposed to the eggs per gram values which would be 0, 24, 48 etc. This factor will vary according to the technique (e.g. Kato-Katz or McMaster), the numbers of slides read per sample, and the value assumed for the mass of feces per slide. This factor was taken or inferred from each article, as detailed in **S1 Table**.

Estimation of negative binomial parameters from grouped data

The data are available as the number of subjects within each category of infection intensity. For example, the World Health Organization defines light, moderate and heavy intensity infections as 1-1,999, 2,000-3,999 and $\geq 4,000$ epg, respectively (Montresor et al. 2002; World Health Organization 2012). However, some publications used different categories. To estimate the two parameters of the negative binomial distribution from such data, the following procedure was used. Within each category, the probability of each possible value in that range occurring was calculated, as a function of the two parameters (μ and k), and summed to give a probability per category. There is no upper bound for the most intense category, so, to avoid an infinite sum, this probability was calculated as 1 minus the sum of the probabilities of the lower categories. The overall log-likelihood was then calculated by multiplying, within each category, the number of observations by the logarithm of the probability, then summing this product over categories.

Finally, the likelihood was maximized over values of μ and k to obtain the fitted values of these parameters. This method requires at least three categories of infection intensity with positive numbers of women. This method was checked by generating simulated data, categorizing them, and ensuring that the original parameter values were recovered.

Results and discussion

Comparison of the negative binomial distribution and the Pareto principle

The pairs of negative binomial mean (μ) and dispersion (k) parameters which comply with the Pareto principle are shown in Fig 1. As the mean increases, the corresponding value of k follows a sawtooth, but generally declining, pattern, towards an asymptote of 0.244. The complexity of this sawtooth pattern is presumably related to the discrete nature of the distribution, and seems to preclude it being expressed exactly by a simple equation.

Fig 1. Relation between the Pareto principle, parameters of the negative binomial distribution, and aggregation of hookworm. The black line represents those combinations of the parameters of the negative binomial distribution which agree with the Pareto principle. This means that, on the black line, 80% of parasites (or any other attribute) will be found in 20% of people. The vertical axis for k , and the horizontal for the mean (μ) are both logarithmic. Each orange dot represents the negative binomial parameters from one study of hookworm in pregnancy: the majority are below the black line, indicating a greater degree of clustering than indicated by Pareto principle.

The Pareto principle (PP) affirms that the Lorenz curve (Butler and McDonald 1989) passes through the 80:20 point. For values of the mean in the range of R_0 for various infections, values of k around 0.3 satisfy the PP (Lloyd-Smith et al. 2005; Susswein and Bansal 2020). Here we have found such k values over a wider range of values of the mean, while taking into account the discrete nature of the NBD. We also sought an explicit expression for the curve in Fig 1, and hence for the value of its asymptote, using generalized gamma, and lognormal, approximations to the negative binomial distribution. For the former, Butler and McDonald (1989) provide an approximation to the normalized incomplete moment and state, moreover, that this itself has another generalized gamma distribution. However, in general, the first moment does not have a closed form solution (Evans et al. 2000). Turning to the lognormal, such an approximation is appealing because, for large μ , the variance of the negative binomial is close to μ^2/k . This means that, for constant k , the negative binomial distribution has constant coefficient of variation (i.e. ratio of the standard deviation to the mean), a property shared with the lognormal distribution. Moreover, the Lorenz curve for the lognormal has a very simple form (Krause 2014). However, equating the first two moments of the two distributions does not give a good approximation, nor does the suggestion of Best and Gipps (Best and Gipps 1974) to equate the second and third moments (not shown). Hence we have not been able to find an explicit expression for the asymptote in Fig 1.

Case study: hookworm in pregnancy

S1 Table shows that the mean (μ) and dispersion (k) parameters could be estimated in 9 of the 13 (70%) component studies in the review, these being the ones with three or more categories of

infection intensity. **S1 Fig** shows that these fitted distributions give a qualitatively good fit to the observed data. However, a limitation of this analysis is the inclusion of one study (Sill et al. 1987) which used a modified Kato technique with Ovo-FEC kits, which is semi-quantitative. The 9 pairs of parameter estimates are shown as points in Fig 1. Seven studies showed super-Pareto variation, i.e. more than 80% of the parasites in 20% of the population (below the line in Fig 1), one was *sub-Pareto* (less than 80% of the parasites in 20% of the population; above the line in Fig 1), and one was *equi-Pareto*. Hence, the parasite distribution in these studies tended to be even more concentrated than implied by the Pareto principle.

The Pareto principle has been proposed as a guide to identifying priority control areas for parasitic diseases (Pan American Health Organization 2019). Although we have found that hookworm infections in pregnant women, at the individual level, are usually super-Pareto, other parasites and insects may be nearly equi-Pareto. For example, this has been reported for ectoparasites, and specifically ticks (Warwick et al. 2016). For disease vectors, a recent study found that bloodfed *Anopheles* mosquitoes per individual were equi- or super-Pareto (Guelbeogo et al. 2018). For phlebotomine vectors of leishmaniasis, a recent study (Moreno et al. 2020) analyzed abundance at five sites in Colombia, using negative binomial models, and the numbers per house were super-Pareto only two of the sites and the other three sites were sub-Pareto. Overall, it is important for control programs to reach those with highest parasite densities or vector exposure, including potential superspreaders (Cooper et al. 2019; Harrison and Bennett 2012).

Conclusions

The case study of hookworm infection in pregnant women did not support the use of the Pareto principle as an empirical description of parasites clustering within people. Rather, in most component studies of the systematic review, the clustering was super-Pareto, i.e. with greater aggregation than stated by the principle.

The negative binomial distribution was the alternative model, and, for this, we developed a method for estimating its parameters from grouped data, in particular the data tables provided in the component studies of the systematic review. We also found, empirically, the relation between those values of the negative binomial distribution's mean and dispersion parameters for which the Pareto principle holds. We were not able to find a closed form expression for the line that embodies these pairs of values, although future work could exploit the relation between the Pareto principle and the Lorenz curve of ecology.

Finally, the existence of super-Pareto aggregation emphasizes the need for control programs of parasitic diseases to reach those few people who harbor a high proportion of the parasites.

Acknowledgments

We appreciate Simon Brooker providing information from the systematic review work on pregnant women. Karim Anaya-Izquierdo helped orient us in our pursuit of an explicit expression for the Lorenz curve. This article includes findings from a master's degree in Epidemiology conducted by Yenifer Orobio at the School of Public Health of the Universidad del Valle in Cali, Colombia, and we thank the tutors and assessors for their guidance during this

degree project. We are also grateful to multiple statistician and epidemiologist colleague, whose ingenious ideas which helped structure the research question and approach, and to the evaluators of the degree project, whose questions helped enrich the discussion.

Declarations

Funding: NA receives salary support from grant number 1U19AI129910-01 from the National Institutes of Health of the United States of America (www.nih.gov) awarded to Dr Nancy Gore Saravia. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Conflicts of interest: The authors have declared that no conflicts of interest exist.

Availability of data and material: All data are included in the manuscript and its supporting information.

Ethics approval: This study was reviewed by the Human Ethics Committee of the Faculty of Health of Universidad del Valle (approval number 083-015).

References

- Anderson RM, May RM (1991) Infectious Diseases of Humans: Dynamics and Control, 1st edn. Oxford University Press, Oxford
- Best DJ, Gipps PG (1974) An improved gamma approximation to the negative binomial. *Technometrics* 16(4):621-624
- Brooker S, Hotez PJ, Bundy DA (2008) Hookworm-Related Anaemia among Pregnant Women: A Systematic Review. *PLoS Negl Trop Dis* 2(9):e291
- Butler RJ, McDonald JB (1989) Using incomplete moments to measure inequality. *Journal of Econometrics* 42(1):109-119

- Cooper L, et al. (2019) Pareto rules for malaria super-spreaders and super-spreading. *Nat Commun* 10(1):3939 doi:10.1038/s41467-019-11861-y
- Evans M, Hastings N, Peacock B (2000) *Statistical Distributions*, third edn. Wiley, New York
- Guelbeogo WM, et al. (2018) Variation in natural exposure to anopheles [sic] mosquitoes and its effects on malaria transmission. *eLife* 7 doi:10.7554/eLife.32625
- Hardy M (2010) Pareto's law. *The Mathematical Intelligencer* 32(3):38-43
- Harrison A, Bennett NC (2012) The importance of the aggregation of ticks on small mammal hosts for the establishment and persistence of tick-borne pathogens: an investigation using the R0 model. *Parasitology* 139(12):1605-13 doi:10.1017/S0031182012000893
- Holland C (ed) (2013) *Ascaris: The Neglected Parasite*. Academic Press, London
- Hotez PJ, Brooker S, Bethony JM, Bottazzi ME, Loukas A, Xiao S (2004) Hookworm infection. *New England Journal of Medicine* 351(8):799-807
- Jankowski MD, Williams CJ, Fair JM, Owen JC (2013) Birds shed RNA-viruses according to the pareto principle. *PLoS One* 8(8):e72611 doi:10.1371/journal.pone.0072611
- Krause M (2014) Parametric Lorenz curves and the modality of the income density function. *Review of Income and Wealth* 60(4):905-929
- Lloyd-Smith JO (2007) Maximum likelihood estimation of the negative binomial dispersion parameter for highly overdispersed data, with applications to infectious diseases. *PLoS One* 2(2):e180 doi:10.1371/journal.pone.0000180
- Lloyd-Smith JO, Schreiber SJ, Kopp PE, Getz WM (2005) Superspreading and the effect of individual variation on disease emergence. *Nature* 438(7066):355-9
- Loukas A, et al. (2016) Hookworm infection. *Nat Rev Dis Primers* 2:16088 doi:10.1038/nrdp.2016.88
- Martínez de la Ossa D, et al. (2010) Parasitosis intestinal. *Ciencia y Salud Virtual* 2(1):122-129
- Montresor A, Crompton DWT, Gyorkos TW, Savioli L (2002) *Helminth Control in School-age Children: A Guide for Managers of Control Programmes*. World Health Organization, Geneva
- Moreno M, Guzmán-Rodríguez L, Valderrama-Ardila C, Alexander N, Ocampo CB (2020) Land use in relation to composition and abundance of phlebotomines (Diptera: Psychodidae) in five foci of domiciliary transmission of cutaneous leishmaniasis in the Andean region of Colombia. *Acta Tropica* 203:105315
- Organización Mundial de la Salud (1994) *Medios auxiliares para el diagnóstico de las parasitosis intestinales*. OMS, Ginebra
- Organización Mundial de la Salud (2005) *Prevención y control de la esquistosomiasis y las geohelmintiasis transmitida por el suelo: informe de un comité de expertos Serie de Informes Técnicos. vol 912*. OMS, Ginebra
- Organización Panamericana de la Salud (2012) *Sobre Geohelmintiasis*. Pan American Health Organization.
https://www.paho.org/col/index.php?option=com_content&view=article&id=1479:sobre-geohelmintiasis&Itemid=561 Accessed 28 December 2020
- Pan American Health Organization (2019) *Stratification of Malaria Based on Risk of Transmission and Elimination of Foci: Region of the Americas. Draft edn*. PAHO, Washington
- Pawlowski ZS, Schad GA, Stott GJ (1992) *Infección y anemia por anquilostomas: posibilidades de prevención y lucha*. Organización Mundial de la Salud, Ginebra

- R Core Team (2019) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria
- Roca C, Balanzó X, Sauca G, Fernández-Roure JL, Boixeda R, Ballesterf M (2003) Uncinariasis importada por inmigrantes africanos: estudio de 285 casos. *Medicina Clínica* 121(4):139-141
- Sill P, Hill A, Igo J (1987) Multifactorial aetiology of anaemia of pregnancy in Port Moresby, Papua New Guinea. *Papua New Guinea Medical Journal* 30(3):193
- Stein RA (2011) Super-spreaders in infectious diseases. *Int J Infect Dis* 15(8):e510-3 doi:10.1016/j.ijid.2010.06.020
- Susswein Z, Bansal S (2020) Characterizing superspreading of SARS-CoV-2 : from mechanism to measurement. medRxiv doi:10.1101/2020.12.08.20246082
- Valerio L, Sabrià M, Fabregat A (2002) Las enfermedades tropicales en el mundo occidental. *Medicina Clínica* 118(13):508-514
- Venables V, Ripley B (2002) *Modern Applied Statistics with S*, 4th edn. Springer Science+Business Media, New York, NY
- Warwick BT, et al. (2016) Abundance estimations of ixodid ticks on Boran cattle and Somali sheep in Northern Tanzania. *International Journal of Acarology* 42(1):12-17
- Woolhouse MEJ, et al. (1997) Heterogeneities in the transmission of infectious agents: Implications for the design of control programmes. *Proceedings of the National Academy of Sciences of the United States of America* 94(1):338-342
- World Health Organization (1992) *Bench aids for the diagnosis of intestinal helminths*. Geneva: World Health Organization
- World Health Organization (2012) *Soil-transmitted helminthiasis: eliminating soil-transmitted helminthiasis as a public health problem in children: progress report 2001-2010 and strategic plan 2011-2020*. World Health Organization, Geneva

