Alexander *Emerging Themes in Epidemiology* (2015) 12:11
DOI 10.1186/s12982-015-0029-4

EMERGING THEMES IN EPIDEMIOLOGY

ANALYTIC PERSPECTIVE Open Access

What's more general than a whole population?

Neal Alexander

Abstract

Statistical inference is commonly said to be inapplicable to complete population studies, such as censuses, due to the absence of sampling variability. Nevertheless, in recent years, studies of whole populations, e.g., all cases of a certain cancer in a given country, have become more common, and often report *p* values and confidence intervals regardless of such concerns. With reference to the social science literature, the current paper explores the circumstances under which statistical inference can be meaningful for such studies. It concludes that its use implicitly requires a target population which is wider than the whole population studied — for example future cases, or a supranational geographic region — and that the validity of such statistical analysis depends on the generalizability of the whole to the target population.

If Czech history could be repeated, we should of course find it desirable to test the other possibility each time and compare the results. Without such an experiment, all considerations of this kind remain a game of hypotheses [1].

Introduction

Classical frequentist statistics relies on the notion of a sampling population to justify probability statements such as *p* values or confidence interval coverage. A sampling population is a source of variation over putative repeated samples, as illustrated in Fig. 1. For such variation to occur, the sample must be smaller than the population, otherwise 'sampling errors disappear altogether' [2] (p659) and *p* values tend to zero. This is formalized in the finite population adjustment, which reduces standard errors towards zero as the size of the sample approaches that of the population [3] (p436). This adjustment is rarely used because, in the classical framework, the sample is much smaller than the population. Sometimes this inequality is simply asserted — e.g., 'We cannot study all the population' [4]. However, advances in computerization of health information, for example through national cancer registries [5] and mass genotyping [6], have made it more feasible to study groups which can reasonably be called 'whole populations'. Although missing data can rarely, if ever, be ruled out, some studies, for example based on cancer registries, have achieved very low levels [7, 8]. The social sciences recognise repeated sampling to be inapplicable to certain kinds of whole population studies [9]. For example, when studying characteristics of the ten largest cities in a given country, based on data aggregated from the latest national census, re-doing the study would not subject the data to sampling variation. To understand how statistical inferential might, nevertheless, be applicable to whole population studies, we need to distinguish different uses of the word 'population'.

One way to preserve sampling-based statistical inference is to adduce a group that is even more general than the whole population that was studied. This kind of wider group is called here a target population, although the term is sometimes used as a synonym for the sampling population [10]. The target population is the group 'about which conclusions are to be made' [11] but, when it differs from the sampling population, it is not subject to the sampling variation formalism. Some authors define target populations to be real [12] (p361) while for others they do not necessarily exist, at least not yet. For example, Kirkwood and Sterne say that enumeration of a target population may be impossible because it 'often includes not only all persons living at present but also those that may be alive at some time in the future' [13]

Correspondence: neal.alexander@lshtm.ac.uk
MRC Tropical Epidemiology Group, London School of Hygiene and Tropical Medicine, Keppel Street, London, UK

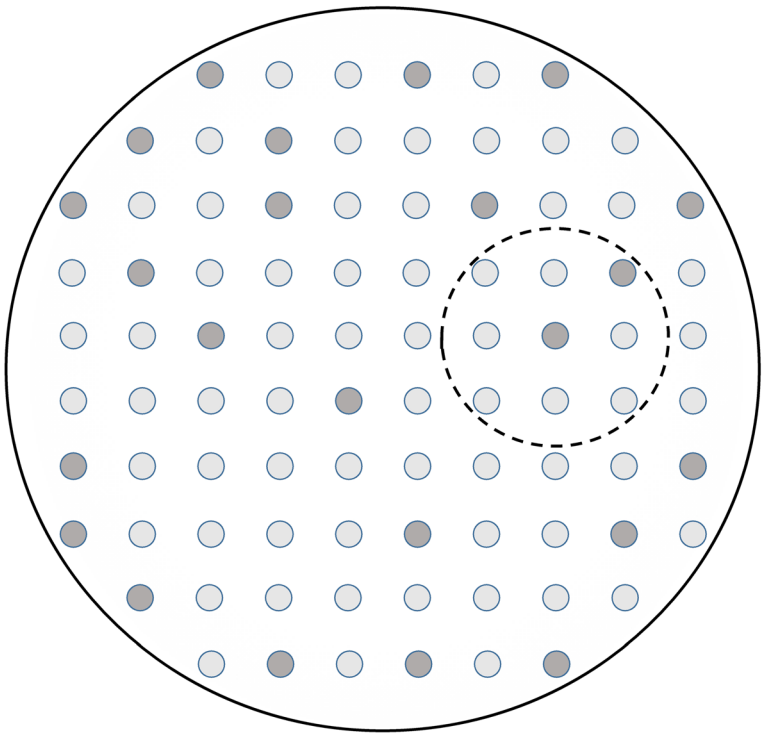

Fig. 1 The classical situation of a sample (small dashed circle) drawn from a population (large solid circle). The darker shading represents units with a characteristic of interest

(p10). Despite being currently intangible, such target populations can be important to the application in hand, the cited authors' example being future recipients of an experimental vaccine. Target populations (in this sense) are sometimes called superpopulations, which comprise 'all possible persons that ever were or ever could be targets of inference' [12].

In the following sections we look at attempts to retain statistical inference in whole population studies, then argue that the problem is better seen as one of generalizability.

Population-free statistics?

Attempts have been made to justify statistical inference to whole populations by resorting to alternative techniques, in particular the bootstrap, and Bayesian estimation [9, 14].

The bootstrap generates variation by sampling the original data with replacement. This can actually be carried out in practice, unlike the notional repeated sampling of a population. However the validity of the bootstrap depends on 'independent identical sampling from an unknown distribution' [15] so the original data are themselves still assumed to result from a sampling process.

Bayesian inference is based on *a priori* probability distributions ('priors') and a likelihood model for data and parameters. The meaning of the prior probabilities does not rely on sampling properties of populations. By applying Bayes' theorem to the model and available data, inference is made about the parameters, for example in the form of credible intervals, i.e., percentile ranges calculated from the parameters' posterior distributions. If we already have complete data on the whole population, then statements about probabilities must refer to something wider. Since we have complete data on the population why do summary statistics not suffice, without confidence or credible intervals? For some situations summary statistics *will* suffice (see Table 1). If they do not, it is because we are interested in a wider group than the original 'whole' population, namely the target population. At the technical level, Bayesian analysis can be run without giving any thought to this target population but this does not mean that the results are validly applicable to it. Moreover, many Bayesian problems allow 'matching' priors which replicate frequentist results [16]. It would seem vacuous to consider one approach valid and the other invalid when they can be designed to give the same results. Similarly, inference, whether Bayesian or not, may also be expressed in terms of parameters of a hypothesised data generating mechanism, often expressed as a probability model. This can be done even when the data comprise a whole population but if the parameters are not estimated perfectly then more could be learned from further data, i.e., from the target population, assuming the same data generating mechanism applies.

Table 1 Consider the following two examples of a binary outcome with complete population coverage

Presidential election: there is 100% turnout in a national presidential election, with each vote being either for candidate A or candidate B.
Cancer registry: male or female sex is registered for all cases in a national registry which has 100% coverage of the type of cancer in question.
In the election example, is it meaningful to estimate a sampling error for the proportion voting for candidate A? The answer seems to be clearly 'no'. This is because the purpose of the election is to choose a president, which is done on the basis of the observed proportion of votes cast. Any kind of interval estimate serves no purpose because there is no generalizability beyond the election.
In the cancer registry example, is it meaningful to estimate a sampling error for the proportion of cases who are female? Some would say 'no' on the basis that it's a complete population enumeration with no sampling error. Similar examples in the literature show that some authors would say 'yes'. This implies an attempt to generalize beyond the population observed, but what is this wider target population? Conceivably future cases, or a wider, supranational geographical area, although often this is left unspecified.

Representativeness and generalizability

The above reasoning suggests that the validity of either frequentist or Bayesian inference depends on the extent to which the group being studied — whether or not it is a whole population — is representative of the target population. Porta [10] allows a sample to be representative of a population if it is 'typical in respect of certain characteristics, however chosen,' i.e., without requiring a particular selection method such as random sampling. Hence a study group can potentially be representative of a given target population, and its findings generalizable to it, even if the target population is not amenable to sampling.

The current paper argues that the problem of inferential methods in whole populations is most usefully understood as one of generalizability (Fig. 2). Thinking in these terms helps show that inferential methods are meaningless for some situations while for others their use is at least arguable (see Table 1).

Lack of information on generalizability has been identified as a limiting factor in the translation of research findings to policy [17–19]. The CONSORT guidelines for clinical trials [20], and STROBE for observational

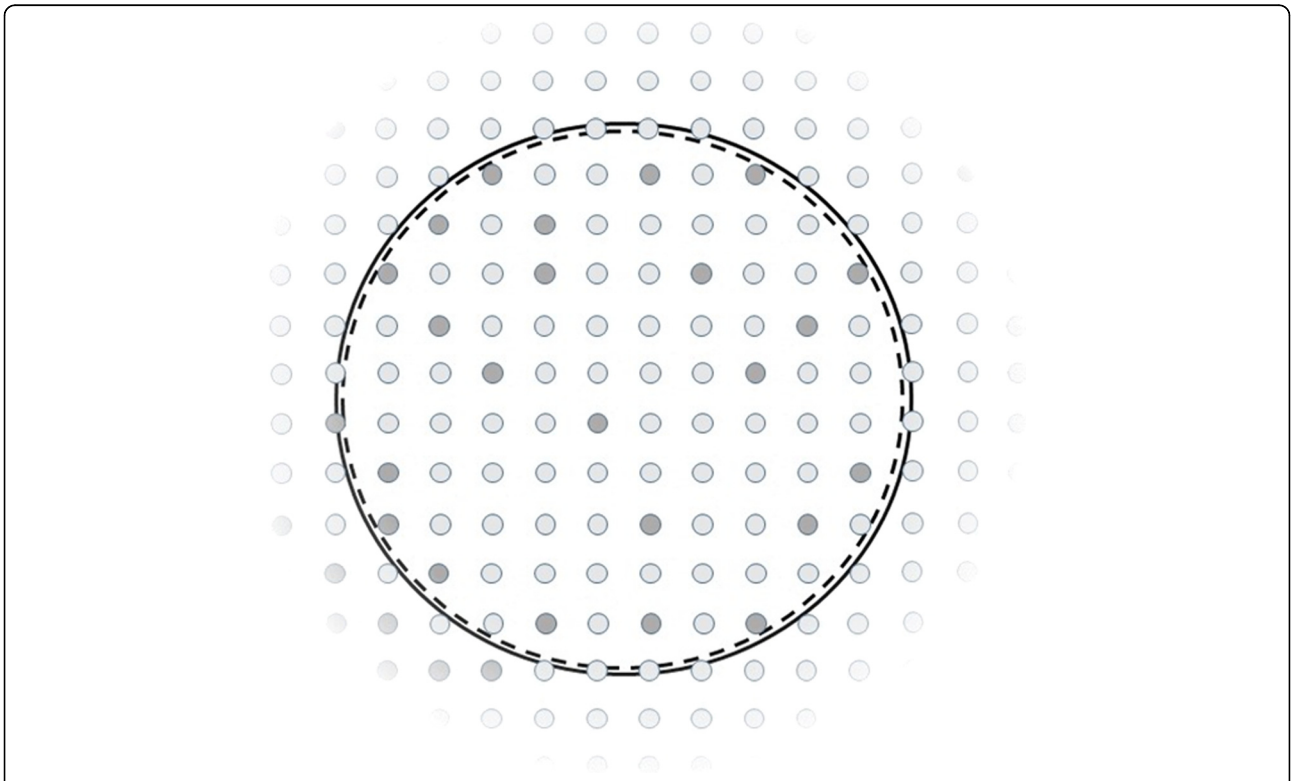

Fig. 2 In a whole population study, the sample (dashed circle) has become so large as to coincide with the population (solid circle). The use of confidence intervals, *p* values, or similar probability statements implies a claim to generalizability to some group beyond the study population: namely the target population, represented by the area outside the two circles. As in Fig. 1, presence or absence of a characteristic is indicated by darker or lighter shading. Authors of whole population studies often do not try to delimit their generalizability. In the figure, such indeterminacy is represented by the target population fading away from the original population: we may not know to what spatial or temporal range the findings may be applicable. Similarly, we may not be able to judge whether the prevalence of the characteristic remains similar in the target population, rather than increasing or decreasing (as it does to the bottom left and top right, respectively, of the figure)

studies [21] mandate discussion of generalizability but give no guidance on how to assess it. Point 21 of STROBE, for example, is 'Discuss the generalisability (external validity) of the study results'. Several frameworks for assessing generalizability have been proposed [17, 22–24]. Inferential analysis of whole population studies is justified only as far as such studies are generalizable, yet reporting of generalizability in other kinds of study is often poor [18, 19]. The extent to which this also applies to whole population studies is briefly assessed in the following section.

Literature review on generalizability of whole population studies

A literature review was carried out to assess the extent to which whole population studies assess their own generalizability. The search term 'whole population [TI]' was used in PubMed on 25 February 2015, restricted to publications in 1994 or later, yielding 64 publications. They were retained if containing primary data on studies of humans, and reporting *p* values or confidence intervals in the abstract. The full text of each of the resulting 13 studies was reviewed for description of its generalizability or external validity (Table 2). The populations were typically either whole countries (e.g., Iceland) or subnational administrative regions of different sizes (e.g., Western Australia, Isle of Wight). Only two papers [25, 26] made reference to wider target populations to which the results might be generalized, and which might give meaning to the *p* values and/or confidence intervals used. These findings can only be suggestive because a limitation of the search is that it does not cover all studies which the current paper would call 'whole population'. For example, the previously-cited Danish National Acute Leukemia Registry [7] has estimated coverage of more than 98 %, and a PubMed search for its name yields 11 studies, but none of these met the current search criteria. Among the studies which *were* identified, however, few are concerned about their generalizability despite the fact that, on the reasoning of the current paper, the validity of their statistical inference depends on it. In particular none of them mentioned the STROBE guidelines which mandate that generalisability be considered.

Table 2

Publication Year	Reference Number	Population	Assessment of generalizability or external validity
2015	[26]	Western Australia	'We are confident about the generalisability of our analytic findings on associations with stimulant medication use Australia-wide'
2014	[33]	Western Australia	none
2014	[25]	North East Scotland	'Our results could be generalisable to young children across the UK'
2014	[34]	Western Australia	none
2013	[35]	Western Australia	none
2012	[36]	Western Australia	none
2012	[37]	Western Australia	none
2012	[38]	Iceland	none
2011	[39]	Western Australia	none
2010	[40]	Scotland	none
2004	[41]	'Top End' of the Northern Territory of Australia	none
2001	[42]	Isle of Wight, United Kingdom	none
1998	[43]	Isle of Wight, United Kingdom	none

Conclusions

Whole population studies often calculate *p* values and confidence intervals which have no explicit theoretical basis. This is a problem that social sciences have, perhaps, confronted more directly than epidemiology [9, 27]. Having concluded that purely statistical workarounds are futile, should we eschew inferential statistics altogether and concentrate instead on descriptive statistics [28]? The current paper advocates a less absolute position; that the conclusions of whole population studies are valid to the extent that their study populations are representative of a wider target population. Hence the problem is converted into one of generalizability. In turn this should be manifest in each study's research question. For example, who is intended to benefit from analysis of national cancer registry data to be applied; future cases in the same country, and/or those further afield? If the former, making this explicit may highlight needs for particular statistical methods, such as time series analysis on the existing data disaggregated by time. Some target populations may not be possible to sample, e.g., because they lie in the future. Similarly, causal inference deals with unobserved or counterfactual outcomes [29, 30], and can be cast in terms of target and source populations [31], and is a promising approach for analysis of whole population studies.

In epidemiology, although aspirations to generalizability are not always met, groundwork has been done on its objective assessment [17, 22–24, 32]. For inferential statistics of whole population studies to be more meaningful, they should fully comply with existing guidelines on reporting generalizability [20, 21], and these guidelines should themselves be updated to reflect the developing best practice.

Competing interests
The author declares that he has no competing interests.

Author contributions
NA conceived the study, carried out the literature search and produced the figures. NA wrote the manuscript and approved the final version.

Acknowledgements
I receive support from the United Kingdom Medical Research Council (MRC) and Department for International Development (DFID) (MR/K012126/1). I am grateful to Karim Anaya-Izquierdo for introducing me to the field of generalizability, and to him, Lyda Osorio, and anonymous referees for helpful discussion of an earlier version of this paper.

Received: 30 June 2014 Accepted: 22 May 2015
Published online: 25 August 2015

References

1. Kundera M. The Unbearable Lightness of Being. London: Faber and Faber; 1984.
2. Armitage P, Berry G, Matthews JNS. Statistical Methods in Medical Research. 4th ed. Oxford: Blackwell Scientific Publications; 2001.
3. Snedecor GW, Cochran WG. Statistical Methods. 8th ed. Ames: Iowa State University Press; 1989.
4. Fathalla MF, Fathalla MMF. A Practical Guide for Health Researchers. World Health Organization Regional Office for the Eastern Mediterranean: Cairo; 2004.
5. Burr JM, Mitry E, Rachet B, Coleman MP. Survival from uveal melanoma in England and Wales 1986 to 2001. Ophthalmic Epidemiol. 2007;14(1):3–8.
6. Hakonarson H, Gulcher JR, Stefansson K. deCODE genetics. Inc Pharmacogenomics. 2003;4(2):209–15.
7. Ostgard LS, Norgaard JM, Severinsen MT, Sengelov H, Friis L, Jensen MK, et al. Data quality in the Danish National Acute Leukemia Registry: a hematological data resource. Clinical epidemiology. 2013;5:335–44.
8. Barlow L, Westergren K, Holmberg L, Talback M. The completeness of the Swedish Cancer Register: a sample survey for year 1998. Acta Oncol. 2009;48(1):27–33.
9. Berk RA, Western B, Weiss RE. Statistical inference for apparent populations. In: Marsden PV, editor. Sociological Methodology 1995. Cambridge: MA: Blackwell Publishers; 1995. p. 421–58.
10. Porta M. Dictionary of Epidemiology. 5th ed. Oxford: Oxford University Press; 2008.
11. Coggon D, Rose G, Barker DJP. Epidemiology for the Uninitiated. 4th ed. London: BMJ Publishing; 1997.
12. Rothman KJ, Greenland S. Modern Epidemiology. 2nd ed. Philadelphia: Lippincott-Raven Publishers; 1998.
13. Kirkwood BR, Sterne JAC. Essentials of Medical Statistics. 2nd ed. Oxford: Blackwell Scientific Publications; 2003.
14. Bollen KA. Apparent and nonapparent significance tests. In: Marsden PV, editor. Sociological Methodology 1995. Cambridge: MA: Blackwell Publishers; 1995. p. 459–68.
15. Efron B. Nonparametric estimates of standard error: the jackknife, the bootstrap and other methods. Biometrika. 1981;68(3):589–99.
16. Bayarri MJ, Berger JO. The interplay of Bayesian and frequentist analysis. Stat Sci. 2004;19(1):58–80.
17. General Accounting Office. Cross design synthesis: a new strategy for medical effectiveness research. Washington, DC: GAO; 1992.
18. Dzewaltowski DA, Estabrooks PA, Klesges LM, Bull S, Glasgow RE. Behavior change intervention research in community settings: how generalizable are the results? Health Promot Int. 2004;19(2):235–45.
19. Klesges LM, Williams NA, Davis KS, Buscemi J, Kitzmann KM. External validity reporting in behavioral treatment of childhood obesity: a systematic review. Am J Prev Med. 2012;42(2):185–92.
20. Schulz KF, Altman DG, Moher D. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. BMC Med. 2010;8:18.
21. von Elm E, Altman DG, Egger M, Pocock SJ, Gotzsche PC, Vandenbroucke JP. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. BMJ. 2007;335(7624):806–8.
22. Bonell C, Oakley A, Hargreaves J, Strange V, Rees R. Assessment of generalisability in trials of health interventions: suggested framework and systematic review. BMJ. 2006;333(7563):346–9.
23. Dekkers OM, von Elm E, Algra A, Romijn JA, Vandenbroucke JP. How to assess the external validity of therapeutic trials: a conceptual approach. Int J Epidemiol. 2010;39(1):89–94.
24. Watts P, Phillips G, Petticrew M, Harden A, Renton A. The influence of environmental factors on the generalisability of public health research evidence: physical activity as a worked example. The international journal of behavioral nutrition and physical activity. 2011;8:128.
25. Smith S, Craig LC, Raja EA, McNeill G, Turner SW. Prevalence and year-on-year trends in childhood thinness in a whole population study. Arch Dis Child. 2014;99(1):58–61.
26. Ghosh M, Holman CD, Preen DB. Exploring parental country of birth differences in the use of psychostimulant medications for ADHD: a whole-population linked data study. Aust N Z J Public Health. 2015;39(1):88–92.
27. Western B, Jackman S. Bayesian inference for comparative research. Am Polit Sci Rev. 1994;88:412–23.
28. Ballinger C. Why inferential statistics are inappropriate for development studies and how the same data can be better used. In: *MPRA*. University Library of Munich, Germany: Paper 29780; 2011.
29. Hubbard AE, Laan MJ. Population intervention models in causal inference. Biometrika. 2008;95(1):35–47.
30. Pearl J. An Introduction to Causal Inference. Int J Biochem. 2010;6(2):7.
31. Höfler M. Causal inference based on counterfactuals. BMC Med Res Methodol. 2005;5:28.
32. Post PN, de Beer H, Guyatt GH. How to generalize efficacy results of randomized trials: recommendations based on a systematic review of possible approaches. J Eval Clin Pract. 2013;19(4):638–43.
33. Malacova E, Kemp A, Hart R, Jama-Alol K, Preen DB. Long-term risk of ectopic pregnancy varies by method of tubal sterilization: a whole-population study. Fertil Steril. 2014;101(3):728–34.
34. Meuleners LB, Fraser ML, Ng J, Morlet N. The impact of first- and second-eye cataract surgery on injurious falls that require hospitalisation: a whole-population study. Age Ageing. 2014;43(3):341–6.
35. Kemp A, Preen DB, Morlet N, Clark A, McAllister IL, Briffa T, et al. Myocardial infarction after intravitreal vascular endothelial growth factor inhibitors: a whole population study. Retina. 2013;33(5):920–7.
36. Clark A, Morlet N, Ng JQ, Preen DB, Semmens JB. Risk for retinal detachment after phacoemulsification: a whole-population study of cataract surgery outcomes. Arch Ophthalmol. 2012;130(7):882–8.
37. Meuleners LB, Hendrie D, Lee AH, Ng JQ, Morlet N. The effectiveness of cataract surgery in reducing motor vehicle crashes: a whole population study using linked data. Ophthalmic Epidemiol. 2012;19(1):23–8.
38. Palsdottir HB, Hardarson S, Petursdottir V, Jonsson A, Jonsson E, Sigurdsson MI, et al. Incidental detection of renal cell carcinoma is an independent prognostic marker: results of a long-term, whole population study. J Urol. 2012;187(1):48–53.
39. Clark A, Morlet N, Ng JQ, Preen DB, Semmens JB. Whole population trends in complications of cataract surgery over 22 years in Western Australia. Ophthalmology. 2011;118(6):1055–61.
40. Langhorne P, Lewsey JD, Jhund PS, Gillies M, Chalmers JW, Redpath A, et al. Estimating the impact of stroke unit care in a whole population: an epidemiological study using routine data. J Neurol Neurosurg Psychiatry. 2010;81(12):1301–5.
41. Currie BJ, Jacups SP, Cheng AC, Fisher DA, Anstey NM, Huffam SE, et al. Melioidosis epidemiology and risk factors from a prospective whole-population study in northern Australia. Trop Med Int Health. 2004;9(11):1167–74.
42. Arshad SH, Tariq SM, Matthews S, Hakim E. Sensitization to common allergens and its association with allergic disorders at age 4 years: a whole population birth cohort study. Pediatrics. 2001;108(2), E33.
43. Tariq SM, Matthews SM, Hakim EA, Stevens M, Arshad SH, Hide DW. The prevalence of and risk factors for atopy in early childhood: a whole population birth cohort study. J Allergy Clin Immunol. 1998;101(5):587–93.