Finite and Infinite Populations in Biological Statistics: Should We Distinguish Them?

Marcin Kozak
Department of Biometry, Warsaw University of Life Sciences
Nowoursynowska 159, 02-787 Warsaw, Poland
Mobile: +48 608 222 059
e-mail: m.kozak@omega.sggw.waw.pl

Abstract: The paper discusses necessity of distinguishing finite and infinite populations in statistical research, with a focus on biological applications. It is shown that statistics for these two types of populations is different and that lack of proper understanding of this issue may do much harm. [The Journal of American Science. 2008;4(1):59-62]. (ISSN: 1545-1003).

Key words: statistics, survey sampling, inference, interpretation.

1. Introduction

There are finite and infinite populations. Both are studied in biological sciences. Surprisingly, there are people, both statisticians and those who apply statistics to interpret results of their research, who are not aware of it. This is a worrying situation because finite and infinite populations are, basically, different and usually should be approached in completely different manners. It is, therefore, important to be aware of it and to be able to distinguish the two statistical philosophies: that related to a finite and to an infinite population.

Whether or not infinite populations exist is a topic for a broad discussion. The main argument of those who do not accept infinite populations is that the universe is finite, so there is no possibility to consider an infinite number of anything. However, for simplicity, in this paper we assume that an "infinite" population approximates the imaginary infinite population when a number of elements of this population is unimaginable. This approximation is necessary to apply classical statistics and interpret properly the results obtained. Hereafter, we take no notice of this approximation and consider finite and infinite populations, keeping, however, this approximation in mind.

There have been not many people who worked on both types of populations. One of them was Jerzy Neyman, a distinguished Polish statistician, whose works on infinite (e.g., related to confidence intervals and hypothesis testing; see, e.g., Neyman and Pearson 1928a, b and 1933a, b, Neyman, 1942) and finite (e.g., related to stratified sampling; Neyman 1934) populations are known among statisticians throughout the world, and whose contribution to statistics of both finite and infinite populations is indisputable. However, one can feel alarmed by the fact that many researchers are not aware of the existence of these two types of populations.

At the very beginning of an investigation, one has to decide whether one considers a finite or infinite population. This decision may sometimes be difficult. It may, and likely will, affect a way of collecting data (the experimental design); data handling at the estimation stage of the study, including a choice of models and statistical methods to be used; and, last but not least, interpretation of the results obtained. Applying classical statistics for a finite population may do more harm than good. Sometimes it may, however, occur to be a reasonable solution, but it must not be a matter of luck, but the reasonable decision of a statistician. Therefore, it is important for a statistician to be aware of which type of population he or she is dealing with.

The aim of this paper is to discuss the importance of distinguishing finite and infinite populations. For the sake of simplicity, let us base on biological studies, and let us keep in mind that the discussion we provide relates also to other studies. Sometimes a decision on the type of population to study may be tricky and have quite an impact on the results and interpretation. I will not discuss statistics for both cases—there are plenty of textbooks and papers on both of them. I will just show that both philosophies are distinct, and why it is so. I hope to convince the reader that this knowledge is important to apply statistics correctly. The paper is directed to those who study, learn or teach statistics of either infinite or finite populations, and to those who apply statistics. Whichever population type a reader deals with, some parts of the paper will provide him/her just basic information, and the remaining should be interesting for him/her. Nevertheless, because most of the classical statistics deals with infinite populations, I will slightly focus on finite populations and show that finite-population studies are quite often in various fields of biology.

2. When do we consider finite and infinite populations?

Whether one should consider a finite or infinite population depends on a study's aims. Based on them, it normally should be obvious which type of population should be defined and which population the interpretation should be linked to. Below, both types of populations are briefly described in relation to biology.

Infinite populations. Classical biological, economic, engineering and other experiments, in which one aims to study a particular process or processes for a specific population or populations, deal with infinite populations. In such studies, inferences and interpretation have a general meaning that helps understand the processes in the

population. A classical agricultural example is a fertilization experiment, whereby one studies whether fertilizer combinations provide similar yield of a crop species. Here, a number of populations considered is equal to a number of fertilizer combinations, and each such a population is infinite in nature. The elements of the populations may be plants of the cultivar or genotype of a species (e.g., rice, wheat, etc.) grown under the specified conditions, fertilizer combinations in our case. However, the populations should be defined more precisely, because the experiments almost always refer to some (broader or narrower) environmental conditions and agrotechnical practices.

There may be a lot of various infinite-population studies in biology, such as studies on resistance of a plant to a pathogen; occurrence of a species in a high-moisture soil environment; or co-existence of two eriophyoid mite species on *Pinus sylvestris* in a National Park. All of such studies have this similarity that they aim at *general* interpretation of the biological processes of study, irrespective of a time point.

Finite populations. Each population the elements of which exist in a particular time is finite. However, some populations are so large that one is unable to imagine all its elements. For example, a number of plants of rice in the world harvested in 2005 is impossible to be counted so may be thought of as an infinite population. On the other hand, a population of farms in India on 1 November 2006 (or November 2006) is imaginable, countable, and as such, should be thought of as a finite population. Note that a finite population may be considered in several occasions; it is still, however, the finite population the elements of which may be changed, removed, or added.

The most common finite-population studies are those related to official statistics. It is enough to mention here agricultural censuses (Wanke 2003) and sample surveys (Kursa and Lednicki 2006) conducted by statistical offices. The censuses/surveys aim to provide information on agricultural production and market in the country and its provinces.

Questionnaire surveys are another example of a finite-population approach. Taking into account only chosen biosciences, they may be useful in surveys on various food science and epidemiology problems (e.g., Willett et al. 1985, Khokhar and Fenwick 1994, Richardson-Harma et al. 1998, Sandström and Faergemann 2004), agricultural practices and production (e.g., Jackson-Smith 1997, Nazarko et al. 2003, Herzog et al. 2006), ecology and wildlife biology (Seiler et al. 2004, White et al. 2005, and the citations therein), not to mention health sciences (e.g., Sammarco et al. 2001). One must be aware, however, that questionnaire samples may also relate to infinite populations (which depends on how a researcher defines a sample and a population). There are many methodological problems connected with questionnaire surveys, like difficulties in obtaining a random sample, high non-response rates, and the like (Särndal et al. 1992).

The following are two finite-population examples from botany/phytosociology. Studying a particular site for occurrence of a particular plant species, a researcher takes samples from this site (Elzinga et al. 1998). The population studied is the site, that is, the geographical area, and the variable of interest—occurrence of the species. A parameter to be estimated may be, for example, a total number of occurrences of the species in the site. If, for any reason, the whole site area may not be surveyed, the samples are taken subject to a chosen sampling design, and the parameter is estimated based on formulas appropriate for the design.

Consider another survey in which one aims to study occurrence of a plant species in National Parks of a country. From each National Park, samples from several randomly chosen sites (the National Park is/may be divided to) are taken to determine whether this species occurs there. This is a stratified sample where a National Park constitutes a stratum from which a sample is taken. A sample from each National Park (stratum) may be taken using simple random sampling without replacement (e.g., quadrat samples from the whole stratum [National Park]—see Elzinga et al. 1998) or more complex designs (especially when the National Park is geographically large) such as a stratified or multistage sampling design. If the complex designs are used within National Parks, the whole design (i.e., for the whole population) becomes complex since it is stratified sampling with different designs within the strata.

Producer groups (such as a group of organic or ornamental farms) may be surveyed to examine their adaptation to the group's requirements or to determine the group's characteristics (e.g., productivity, agrotechnical practices, environmental conditions, etc.). An animal species in a zoological garden(s) may be studied for infection with a pathogen. Apples in an orchard may be studied for maggoty. Deer in a district may be studied for disease infection. Animals in a country may be studied for disease symptoms after a national epidemic. A much longer list of examples of finite-population biology research might be provided here, but the above-given show the importance of ability to distinguish finite and infinite populations in biology.

3. Why is it important to distinguish finite and infinite populations?

In previous sections, differences between finite and infinite populations have been pointed out and the examples of both types of populations have been given. Now it is time to show why a researcher should be able to distinguish finite and infinite populations.

Classical, infinite-population statistics assumes that a sample be simple, which, basically, means that all the sample observations should be stochastically independent and follow the same theoretical distribution. In

case of sampling from a finite population, a sample is simple only for simple random sampling with replacement, which, practically, is not the case in any survey and has only theoretical importance. Note that limiting distributions in a classical sense are concerned with a simple random sample. These limiting distributions, proper for infinite populations, had to be adapted to work for finite populations (see, e.g., Hájek 1960). This is mainly because samples from finite populations are taken using complex sampling designs and schemes, which makes the samples non-simple. Not to mention more complex sampling designs, such as stratified or multistage sampling, even a simple random sample taken without replacement is non-simple because observations for the population units sampled are dependent. For most finite-population sampling schemes, estimators for simple population parameters, like mean or proportion, are different from those for simple random sample (which work for infinite populations). In addition, formulas for variance of the estimators differ for the two cases, which, in turn, makes the confidence intervals for the estimators also differ. Testing hypothesis for finite populations under non-simple sample is also different from that for infinite populations, and even classical tests need to be modified. Therefore, any classical statistical methodology that is correct for infinite populations is usually not correct for finite populations.

What has already been said relates to statistical theory and application. However, statistics aims to provide tools for researchers to interpret the results of their studies. And here mixing up finite and infinite populations may cause most harm. This is because interpretation about finite populations is based on completely different philosophy and has completely different meaning than that about infinite populations. For finite populations, one usually attempts to picture processes one wants to study in this population, usually in relation to a particular time. This makes the interpretation has no general meaning (as it usually is for infinite populations), but simply refers to this particular set of elements in this particular time. For infinite populations, one may provide interpretation having a general meaning, which is not generally linked to particular elements of the population and to a particular time point. Knowledge of processes in an infinite population may, for example, help the researcher understand some general processes in a population or several populations (e.g., yielding of a crop species in particular environmental conditions; resistance of a cultivar to a pathogen; influence of fertilization on a yielding level; occurrence of a species in a high-moisture soil environment; co-existence of two eriophyoid mite species on *Pinus sylvestris* in a National Park in general; etc.). Inferences based on finite populations are of different character, and usually describe the population in a particular time (e.g., average yielding of a crop species in a country in 2006; resistance of a cultivar to a pathogen in the particular region of a country; influence of fertilization on a yielding level in a population of farms in a region in 2006; occurrence of a species in a National Park in 2006; co-existence of two eriophyoid mite species on *Pinus sylvestris* in a National Park in 2006; etc.).

4. Conclusion

Distinguishing finite and infinite populations is of importance for biostatisticians because they have to choose appropriate methods to apply, and for researchers applying statistics because they have to know how to draw inferences about their populations. Without this basic knowledge, results and interpretation of a research may occur to be false and unreliable, and convincing the researcher that one should consider finite-population or infinite-population approach might be difficult. It is also possible to find a statistician who makes light of an approach he or she does not work on. This is alarming because none of the approaches is better than another: they both are important in various sciences, including biology, and they should not be mixed up.

References

1. Elzinga, C.L., Salzer, D.W., and Willoughby, J.W. Measuring and monitoring plant populations, BLM Technical Reference 1730-1, BLM/RS/ST-98/005+1730. Bureau of Land Management, Denver, Colorado;1998.
2. Hájek, J. Limiting distributions in simple random sampling from a finite population. Publications of the Mathematics Institute of Hungarian Academy of Science 1960;5:361-375.
3. Herzog, F., Steiner, B., Bailey, D., Baudry, J., Billeter, R., Bukacek, R., De Blust, G., De Cock, R., Dirksen, J., Dormann, C.F., De Filippi, R., Frossard, E., Liira, J., Schmidt, T., Stockli, R., Thenail, C., van Wingerden, W., and Bugte, R. Assessing the intensity of temperate European agriculture at the landscape scale. Europ. J. Agronomy 2006;24:165-181.
4. Jackson-Smith, D., Nevius, M., and Bradford, B. Manure management in Wisconsin: Results of the 1995 Wisconsin farmer poll with questionnaire. PATS Res. Rep. 1, Progr. on Agric. Technol. Studies, Univ. of Wisconsin College of Agric. and Life Sci., Madison; 1997.
5. Khokhar, S., and Fenwick, G.R. Phytate content of Indian foods and intakes by vegetarian Indians of Hisar Region, Haryana State. J. Agric. Food Chem. 1994;42:2440-2444.
6. Kursa, L., and Lednicki, B. The agricultural sample surveys in Poland in transition period. Statistics in Transition 2006;7(5):981-1008.
7. Nazarko, O.M., Van Acker, R.C., Entz, M.H., Schoofs, A., and Martens, G. Pesticide Free Production of

Field Crops. Results of an On-Farm Pilot Project. Agron. J. 2003;95:1262-1273.
8. Neyman, J. On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. J. Royal Stat. Soc. Ser. A 1934;97:558-606.
9. Neyman, J. Basic ideas and theory of testing statistical hypotheses. J. Royal Stat. Soc. 1942;105:292-327.
10. Neyman, J., and Pearson, E. S. On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. Biometrika 1928a;20A:175-240.
11. Neyman, J., and Pearson, E. S. On the use and interpretation of certain test criteria for purposes of statistical inference: Part II. Biometrika 1928b;20A:263-294.
12. Neyman, J., and Pearson, E. S. On the problem of the most efficient tests of statistical inference. Biometrika 1933a;20A:175-240.
13. Neyman, J., and Pearson, E. S. The testing of statistical hypotheses in relation to probabilities a priori. Proc. Cambridge Philos. Soc. 1933b;29:492–510.
14. Richardson-Harman, N., Phelps, T., Mooney, P., and Ball, R. Consumer perceptions of fruit production technologies. New Zeal. J. Crop Hort. Sci. 1998;26:181-192.
15. Sammarco, M.L., Ripabelli, G., Grasso G.M. Evaluation of the applicability of a questionnaire on occupational health risks in agriculture. Annali di igiene 2001;13(5):451-61 (in Italian).
16. Sandström, M.H., and Faergemann, J. Prognosis and prognostic factors in adult patients with atopic dermatitis: a long-term follow-up questionnaire study. British Journal of Dermatology 2004;150(1):103-110.
17. Särndal, C.E., Swensson, B., and Wretman, J. Model Assisted Survey Sampling, Springer-Verlag, New York;1992.
18. Seiler, A., Helldin, J.O., and Seiler, C. Road mortality in Swedish mammals: results of a drivers' questionnaire. Wildl. Biol. 2004;10:183-191.
19. Wanke, H. The Censuses of Agriculture in Poland in 1996 and 2002, Statistics in Transition 2003;6(2):275-286.
20. White, P.C.L., Jennings, N.V., Renwick, A.R., and Barker, N.H.L. Questionnaires in ecology: a review of past use and recommendations for best practice. J. Appl. Ecology 2005;42:421-430.
21. Willett, W.C., Sampson, L., Stampfer, M.J., Rosner, B., Bain, C., Witschi, J., Hennekens, C.H., and Speizer, F.E. Reproducibility and validity of a semiquantitative food frequency questionnaire. Amer. J. Epidemiology 1985;122:51-65.