Economics and Human Biology 9 (2011) 419–437

Contents lists available at ScienceDirect

Economics and Human Biology

journal homepage: http://www.elsevier.com/locate/ehb

Comparing population distributions from bin-aggregated sample data: An application to historical height data from France

Jean-Yves Duclos[a,b,1], Josée Leblanc[c], David E. Sahn[d,*]

[a] Institut d'Anàlisi Econòmica (CSIC), Barcelona, Spain
[b] Département d'économique and CIRPÉE, Université Laval, Quebec, Canada G1A OV6
[c] Department of Finance, Ottawa, Canada
[d] Economics, Cornell University, Ithaca, NY 14853, USA

ARTICLE INFO

Article history:
Received 27 May 2010
Received in revised form 3 May 2011
Accepted 3 May 2011
Available online 13 May 2011

JEL classification:
N33
I10
C46

Keywords:
Health and nutrition
Distributional analysis
France
Economic history
Height
Physical stature
Anthropometric history

ABSTRACT

We develop a methodology to estimate underlying (continuous) population distributions from bin-aggregated sample data through the estimation of the parameters of mixtures of distributions that allow for maximal parametric flexibility. The statistical approach we develop enables comparisons of the full distributions of height data from potential army conscripts across France's 88 departments for most of the nineteenth century. These comparisons are made by testing for differences-of-means stochastic dominance. Corrections for possible measurement errors are also devised by taking advantage of the richness of the data sets. Our methodology is of interest to researchers working on bin-aggregated or histogram-type data, something that is still widely done since much of the information that is publicly available is in that form, often due to restrictions based on confidentiality concerns.

1. Introduction

Historical data on stature, a measure of health and nutritional status, are widely available from many countries and have been employed to gain insight into the level and long-term evolution of well-being and living standards, particularly in Europe and the United States. The widespread use of height data (standardized for age and gender) reflects that such anthropometric data is the outcome of a combination of inputs that affect nutrition and disease, such as the local health environment, access to clean water, nutrient intake, maternal health status, health technology, and the organization of work. In short, stature captures "multiple dimensions of the individual health and development and their socio-economic and environmental determinants" (Beaton et al., 1990)[1] and is considered to be the best general indicator of health in a population (de Onis et al., 2000). And in particular, heights of young men

* Corresponding author. Tel.: +1 607 255 8931; fax: +1 607 255 0187.
E-mail addresses: jyves@ecn.ulaval.ca (J.-Y. Duclos), Leblanc.Josee@fin.gc.ca (J. Leblanc), david.sahn@cornell.edu (D.E. Sahn).
[1] Tel.: +1 418 656 7096; fax: +1 418 656 7798.

[1] This is also supported by the existence of a significant correlation between heights and various indicators of health – see for instance Wagstaff (2002) – suggesting that the choice of an indicator other than height would yield similar results. Moreover, adult stature is not only a good indicator of prior episodes of infection and chronic disease, but it has also been shown to be an important determinant of risk of morbidity and mortality.

1570-677X/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.ehb.2011.05.002

entering adulthood are a cumulative indicator of their overall health and nutritional status during their formative years, particularly the period prior to the beginning of puberty.[2]

Economic historians have expended considerable effort to examine changes in anthropometric outcomes of various populations over time (Fogel et al., 1982; Steckel and Floud, 1997; Weir, 1997; Komlos, 1998; Komlos and Baur, 2004; Deaton and Paxson, 1998; Goesling and Firebaugh, 2004).[3] There is strong evidence that heights increased substantially throughout Europe during the 19th century (Steckel, 2006), including in France, which is the focus of this paper (Weir, 1997; Postel-Vinay and Sahn, 2010; Komlos et al., 2003). An important concern that arises in how historians use anthropometric indicators, however, is whether the summary health statistics they employ, particularly measures of central tendencies, can adequately capture the *distribution* of health. It is increasingly recognized that examining entire distributions of health can provide valuable information that would otherwise be hidden by summary health statistics, such as means and the share of the population that falls before a normative standard, or cut-off point, that may define poor health.[4]

We follow the expanding literature that uses historical height data, as well as height data on contemporary malnourished populations, to learn about health inequality by examining the entire distributions of health (Guntupalli and Baten, 2006; Sahn, 2005). Of particular importance to our investigation is that, as pointed out by Deaton (2008), distributions in adult height capture the impact of differences in nutrition and health care that these cohorts experienced as children.[5] Populations with poor nutrition and health conditions are characterized by both a mean shift of heights to the left, and increased skewing of the distribution. Thus, the extent of poor health, as captured by heights falling below a normative threshold, can be decomposed into the contribution of the role of inequality and shifts in the mean (Sahn, 2005; Sahn and Younger, 2005, 2006; Pradhan et al., 2003).

This paper gives prominence to the distributional analysis of health by examining both the distribution of heights throughout France and its evolution in the nineteenth century. More specifically, the paper uses a particularly rich data set collected on men who were called up for possible conscription into the French army during this period. The screening of all men at the age of 20 for mandatory military service involved a physical examination, including measuring their height. The data are not truncated from below (a common problem with height data of this sort), since the samples are of potential conscripts, not of those who were selected.[6] Data in such abundance for a whole century are a rare find. Consequently, socio-economic improvements as well as periods of adverse conditions in 19th-century France can be expected to have an observable effect measured by the stature achieved at 20 years of age.[7]

While individual heights were recorded at the time of conscription, the data we have available are limited to the number of men that fall into classes, or bins of height intervals, for each year and department. This raises obvious challenges given our objective to compare the underlying continuous distributions of heights. These difficulties are not unique to our data, or the paper's historical period of interest. Even today, much of the information that is used and publicly available on distributions of income come from bin-aggregated, or histogram-type, data – important examples of this are the popular World Income Inequality and POVCAL databases.[8] In several countries, the unit data from early household surveys have not survived; in other cases, access to distant and/or more recent microdata is restricted by political sensitivity and confidentiality concerns. We therefore propose and implement a method that estimates the underlying continuous population distributions from the bin-aggregated sample data through the estimation of the parameters of mixtures of distributions that allow for maximal parametric flexibility.[9] This approach will be of general interest to both historians and other researchers working on contemporary bin level data on household incomes, heights, and other indicators of well-being.

[2] Comparing heights of adults (unlike children) across ethnicities is of potential concern. However, this is not a serious problem in our case because our population is relatively homogeneous ethnically.

[3] A number of other studies have also looked at health inequalities, using statistics such as covariances (such as Deaton and Paxson (1998) and Anson and Sun (2004)), differences between various percentiles (Costa, 1999), distances from the mean for different social classes (Anson and Sun, 2004), or inequality indices (e.g., Wagstaff, 2002; Pradhan et al., 2003; Goesling and Firebaugh, 2004; Sahn and Younger, 2005). Studies of health inequality have also attempted to explain whether inequality in some factors, such as income, can transmit itself into health inequality – see for instance Weir (1997), Anson and Sun (2004) and Nolte and McKee (2004). Other studies have decomposed the evolution of health inequality across factors, such as Pradhan et al. (2003), Goesling and Firebaugh (2004), Sahn and Younger (2005), and Wagstaff (2002).

[4] For instance, see Pradhan et al. (2003) and Deaton (2003, 2008).

[5] Deaton also notes in the context of his work on India that owing to the theory of sexual dimorphism, heights will be greater where the ratio of females to males is lowest. This is unlikely to be an important explanation for height inequality in our sample. Nonetheless, we cannot rule out this as a potential contribution to overall inequality, or changes therein over time.

[6] See A'Hearn (2004) for an example of working with a truncated sample.

[7] The age of conscription was constant at 20 years throughout the period which our data spans (Postel-Vinay and Sahn, 2010; Weir, 1997; Baker, 1998). As Major-General Balfour remarked in his lengthy paper on the history of French conscription, the early laws of France dating back to the beginning of the 19th century stipulated ''that every Frenchman owes military service to the State, and shall, therefore, on the completion of 20 years of age, be inscribed on the drawing list, and take his chance of being drawn for the conscription'' (Major-General Balfour, 1867, p. 219). A'Hearn et al. (2009) present statistical corrections to address changes in the age of measurement of Italian conscripts.

[8] See http://www.wider.unu.edu/research/Database/en_GB/database/ and www.worldbank.org/LSMS/tools/povcal/.

[9] See for instance Kleiber and Kotz (2003) for a review of some of the more restrictive functional forms that have been proposed in the literature and further references to the literature. The procedures used in the paper combine features of the literature on the fitting of finite mixtures – see Titterington et al. (1995) and Paapaa and van Dijk (1998) for instance – and of the literature on the estimation of distributions using grouped data – such as in McDonald (1984).