European
Historical
Economics
Society

EHES Working Paper | No. 142 | December 2018

Testing for normality in truncated anthropometric samples

Antonio Fidalgo
HS-Fresenius University of Applied Sciences

EHES Working Paper | No. 142 | December 2018

Testing for normality in truncated anthropometric samples

Antonio Fidalgo[‡]
HS-Fresenius University of Applied Sciences

Abstract

Anthropometric historical analysis depends on the assumption that human characteristics—such as height—are normally distributed. I propose and evaluate a metric entropy, based on nonparametrically estimated densities, as a statistic for a consistent test of normality. My first test applies to full distributions for which other tests already exist and performs similarly. A modified version applies to truncated samples for which no test has been previously devised. This second test exhibits correct size and high power against standard alternatives. In contrast to the distributional prior of Floud et al. (1990), the test rejects normality in large parts of their sample; the remaining data reveal a downward trend in height, not upward as they argue.

JEL Codes: C12, C14, N3, N13, J11

Keywords: test of normality, truncated samples, anthropometrics

‡ HS-Fresenius University of Applied Sciences, Cologne, Germany.
fidalgoantonio@gmail.com. I thank my PhD supervisor, Liam Brunt, as well as John Komlos and Jeff Racine for useful discussion and comments on specific issues. Hamid Hussain-Khan offered kind technical support with the cluster at the University of Lausanne, Switzerland. The usual disclaimer applies.

Notice
The material presented in the EHES Working Paper Series is property of the author(s) and should be quoted as such. The views expressed in this Paper are those of the author(s) and do not necessarily represent the views of the EHES or its members

1 Introduction

Anthropometric history mines archival data on individuals' physical characteristics such as their height in order to assess the evolution of living standards in various contexts (Steckel, 1995). As long recognized by many scholars (Fogel et al., 1983, Floud et al., 1990, Komlos, 1994), such anthropometric indicators have the potential to supplement conventional economic indicators in disputed issues—for instance, the optimists versus pessimists debate on workers' fortunes at the early stages of and during the Industrial Revolution (Lindert and Williamson, 1983, Mokyr, 1988, Komlos, 1998, Nicholas and Steckel, 1991). Anthropometric data may even replace standard indicators when they are either lacking (Cameron, 2003) or unreliable.

As a distinctive feature, this approach has inherited a fundamental prior: within any given homogeneous population, the studied characteristic of human organisms ought to be normally distributed (Tanner et al., 1966).

It follows that all the statistical methods devised to extract information from the available samples—such as maximum-likelihood estimation—rely precisely on the normality assumption (Wachter and Trussell, 1982, Komlos and Kim, 1990, A'Hearn, 2004). But this comes at a cost. If normality is not verified in the analysed sample, then inference from that sample using these methods becomes invalid.

The usual antidote consists of starting the analysis by applying a battery of tests of normality to the sample. Unfortunately, this safe strategy is generally not an option in historical anthropometrics. The reason lies in the nature of the samples commonly available in the field. Coming from military institutions that imposed a minimum height, these samples are left-truncated—so-called deficient samples—because only heights of recruits above the minimum height requirement were recorded.[1] The risk of invalid inference therefore remains acute since, as Komlos (2004) reminds "statistical tests of normality have not been devised for distributions with height requirements".

Consider then the path-breaking contribution in historical anthropometrics by Floud et al. (1990), along with countless subsequent studies, that have drawn conclusions about the secular trend of individuals heights from deficient samples. It appears that they have engaged in standard, yet unsafe, statistical practice whose results are shaded by possible invalid inference.

This paper introduces and evaluates two new tests of normality for full and deficient samples—i.e., a challenger to the existing tests and the first test of its kind for normality in truncated distributions. Both belong to a class of tests based on a metric entropy computed from nonparametrically estimated densities (Granger et al., 2004, Li and Racine, 2007, Racine, 2012). It is shown that their performance is outstanding in simulated data.

In Section 5 below, I re-analyse the Floud et al. (1990) samples in the light of the test developed here. I show that normality can be rejected in large parts of their sample, in particular for those of the youngest recruits. The consequence

[1] In some circumstances, however, the truncation might not have been perfect and some shorter soldiers were allowed to enter.

of these tests is quite dramatic. The upward secular trend sketched by the Floud et al. (1990) estimates turns into a downward secular trend if one restricts calculations to validly inferred estimated heights.

The implication of this result should be seen in the light of the aforementioned controversy on the salutary versus detrimental early effects of the Industrial Revolution on the working classes. As both sides of the debate have built on various indicators, Floud et al. (1990) has certainly served as the main anthropometrical caution of the optimistic side. One that this paper reveals as highly misleading.

Section 2 briefly outlines the pivotal importance of tests of normality in anthropometrics and Section 3 provides details of the tests developed here. Section 4 presents the results of the simulations I have run to evaluate the size and the power of the tests. Section 5 re-analyses the Floud et al. (1990) data. Section 6 offers concluding remarks.

2 Testing for normality in anthropometrics

The issue of distinguishing normal samples from non-normal ones is not new in anthropometrics. Galton (1875), who refers to the "law of frequency of error", matches the empirical distribution of heights to the corresponding percentiles of the normal distribution and judges the normality assumption "fairly applicable" for his sample. (Pearson, 1895) develops a test based on the moments of a distribution and compare them with their counterpart of the normal distribution. His examined samples include height distributions. Over time, this issue has received a considerable amount of attention in the broader context of tests of normality (Pearson et al., 1977, Shapiro et al., 1968, Jarque and Bera, 1987, to cite only a few).

There seems to exist even a general contentment with the available tools—as attested by the following subtle reversal of the chain of implications.[2] Normality of population height implies normality of randomly selected height samples. Hence, some authors (e.g., Nicholas and Steckel, 1991) read failure to reject normality of their sample as evidence of its "cleanliness" and therefore dismiss any possible source of bias such as sample selection or truncation. These authors seem to be confusing necessity and sufficiency: normality of the sample is necessary for it to be randomly drawn from the population, but it is not sufficient to prove it.

Notice, however, that the usual tests of normality are typically inconsistent (Bierens, 1982). A test is called consistent if Prob(Reject $H_0 \mid H_0$ is false) $\to 1$ as $n \to \infty$. Since the power of a test is defined as Prob(Reject $H_0 \mid H_0$ is false), a consistent test has therefore asymptotic power equal to one. The importance of this property is better assessed once contrasted with the traditional tests. The power of these latter depends on the set of alternatives H_1 chosen. In short, the "inconsistency" of those tests arises because the set of H_1's is not a complement of H_0. This means that they may show good power against some

[2] A similar argument about the interpretation of the non-rejection of normality can be found in Bodenhorn et al. (2012).

alternatives but lack thereof against an undetermined set of others (see Yazici and Yolacan, 2007, Noughabi and Arghami, 2011, for a few simulations exhibiting insufficient power). Bodenhorn et al. (2012) have precisely suggested that among these undetectable departures one can count some that are extremely relevant for the anthropometric literature, namely samples plagued by sample selection problems (see also Mokyr and Ó Gráda, 1996). Their conclusion, based on simulations, calls for caution. If samples were selected in the way they conjecture, i.e. one that results in a non-normal distribution of the heights due to the sampling process, then the available tests would not be able to distinguish them from normally distributed samples.

In contrast, the first test proposed here is a consistent test of normality. It makes no assumption on the underlying distribution of the analysed sample and its unknown univariate density is first estimated nonparametrically by a kernel method thanks to a recent implementation in R (Li and Racine, 2007, Racine, 2012, R Core Team, 2012). The test then computes a metric entropy—normalized Hellinger distance of Granger et al. (2004)—for testing the null hypothesis of equality of the estimated density and a normal density with identical first two moments.

I first show that, even in small samples, the test exhibits correct size—around 5% of the truly normal samples are rejected by the 5% nominal test. Then, for a set of usual alternatives—e.g., various t-distributions, uniform distribution, etc.—I find that its power is comparable to the power of the common parametric tests used in the literature (Jarque-Bera, Shapiro-Wilk, d'Agostino). The same applies for polluted distributions such as mixes of normals. For the "quite intractable"[3] cases of sample selection, the results depend on the form of modelling the selection process. In the simple form of sample selection analysed here, the performance of this test is again in line with its competitors' performance. Therefore, I argue that this new test deserves a privileged place in the researcher's statistical toolbox.

The second test proposed here is a modified version of the previous one. As mentioned above, it is prompted by the nature of most historical samples analysed in the heights literature—i.e., military height measurements. Institutions that provided this type of data typically imposed a minimum height restriction for their recruits, *de facto* eliminating the lower part of the height distribution. This practice represents a serious hurdle for statistical analysis and has, consequently, spurred contributions to resolve the estimation problem. Later, I discuss the problem of empirically estimating truncation points. For now, I confine the discussion to the statistical problem of valid inference. [4]

Note first that the standard approach in the height literature depends yet

[3] Komlos (2004, footnote 44).

[4] Restricting to methods dealing with univariate distributions, a fair list—shared with Jacobs et al. (2008)—of the most important ones includes: the reduced-sample/truncated maximum likelihood estimator and the quantile bend estimator (Wachter and Trussell, 1982), the Komlos-Kim method (Komlos and Kim, 1990) and the restricted maximum likelihood estimator (A'Hearn, 2004). Notice that deciding between them is a difficult and probably case specific choice. But it is certainly not an innocent one as reminded by fierce controversies (Komlos, 1993, Floud et al., 1993).

more strongly on the the normal distribution prior. With deficient samples, the validity of that assumption is no longer tested, as it is for the available full distributions. Instead, statistical inference is conducted by imposing the normality assumption on the estimators. Again, this constitutes a potentially problematic departure from sound statistical practice and therefore casts serious doubt on the results obtained with these tools (see Jacobs et al., 2008, for a unique example of the extent of the problem). The approach can be justified, though, on the basis that there is no normality test which applies to a part of a distribution only.

The procedure of the previous test can be adapted to provide a feasible test for deficient samples. The density of the sample truncated at value ξ—typically, the minimum height requirement—is again estimated nonparametrically. The test then computes a metric entropy—normalized Hellinger of Granger et al. (2004)— for testing the null hypothesis of equality of the estimated density and the density of a normal distribution truncated at ξ and with appropriately chosen first two moments. I assume ξ has been correctly identified,[5] even though this is not always easily granted in practice (Komlos, 2004); more on his later.

The size of the test is evaluated for different values of ξ. Overall, the test presents correct size even for relatively small samples. Power experiments are also run for a set of alternatives—$t(5)$, $\Gamma(5,1)$, etc.—and different ξ's. Simulations show good results with almost unit power often reached for samples of 500 observations only. Being alone in its class, this test should prove to be useful for statistical analysis, particularly in the heights literature.

3 An entropy-based test of normality

3.1 The entropy measure

The essence of the test considered here is to evaluate the "similarity" between the densities of two continuous, univariate random variables. Formally, the test is based on the metric entropy—normalized Hellinger of Granger et al. (2004)—given by

$$S_\rho \;=\; \frac{1}{2}\int \left(f^{1/2} - g^{1/2}\right)^2 dx, \tag{1}$$

where f and g are the corresponding marginal densities of the two random variables. H_0 is equality of the two densities. Hence, it will be rejected whenever the two densities are too "distant" from each other—in an entropy sense.

A few words on the merits of this measure and its relation to the entropy *divergence* measures are the following (see Maasoumi, 1993, Ullah, 1996, Granger et al., 2004). Consider the generalized β-class of entropy measures proposed by Havrda and Charvát (1967)

$$H_\beta(f) = \begin{cases} \frac{1}{\beta-1}(1 - Ef^{\beta-1}) & \text{for } \beta \neq 1, \beta > 0, \\ -E\log f & \text{for } \beta = 1, \end{cases} \tag{2}$$

[5] Or above the true minimum height requirement.

where E indicates the expectation with respect to the distribution f. Notice that for the special case $\beta = 1$, this measure reduces to the well-known Shannon's entropy.

Based on these entropy measures, one can define, for any two density functions f and g, the β-class entropy *divergence* of g from f given by

$$H_\beta(f,g) = \frac{1}{\beta - 1} \int f \left[\left(\frac{f}{g} \right)^{\beta - 1} - 1 \right] dx, \quad \beta \neq 1. \tag{3}$$

$H_\beta(f,g)$ is not a symmetric measure.[6] To overcome the symmetry obstacle, one can combine the two asymmetric divergence measures to obtain a symmetric β-class measure. The following option has been proposed to achieve it:

$$I_\beta(f,g) = H_\beta(f,g) + H_\beta(g,f). \tag{4}$$

Notice again that, for the special case $\beta = 1$, this class yields a familiar concept—i.e., the Jeffreys-Kullback-Leibler divergence. As will become clear, however, the interesting value for the current purpose is $\beta = 1/2$. Indeed, for that value, the measure presented here satisfies the triangular distance inequality—hence it is a proper metric. To see this, write

$$\begin{aligned} I_{1/2}(f,g) &= 4 \left[1 - \int (fg)^{1/2} dx \right] = 2 \left[\int (f^{1/2} - g^{1/2})^{1/2} dx \right] \\ &= 2 d_{(2)}(f^{1/2}, g^{1/2}). \end{aligned} \tag{5}$$

where $d_{(2)}(\cdot)$ is the L_2-norm distance between $f^{1/2}$ and $g^{1/2}$ satisfying the triangular inequality. Expression (5) also involves two special *distance* measures—$M(\cdot)$, known as Matusita or order-1 Hellinger distance given by

$$M(f,g) = \int (f^{1/2} - g^{1/2})^{1/2} dx, \tag{6}$$

and $B(\cdot) = 1 - \rho^\star$, with

$$0 \leq \rho^\star = \int (fg)^{1/2} dx \leq 1, \tag{7}$$

where $\rho^\star$ is the Battacharyya coefficient that can be interpreted as a measure of "affinity" between f and g.

To make more explicit the key characteristic of the measure suggested in this paper—i.e., that it is an entropy metric—I redundantly write the following relationships between the measures above

$$S_\rho = 1 - \rho^\star = B = \frac{1}{2} M = \frac{1}{2} d_{(2)} = \frac{1}{4} I_{1/2}. \tag{8}$$

It must be emphasized that, because of the triangle inequality property, $M(\cdot)$ and $B(\cdot)$ are unique in their class of measures of divergence—this, in turn, translates into many advantages for $M(\cdot)$ and $B(\cdot)$ over these latter, notably when applied to non-nested models. For further properties of the entropy measure S_ρ, see Granger et al. (2004) and the references therein.

[6] Obviously, neither is $H_\beta(g,f)$.

3.2 Two special cases for testing normality

The entropy metric described above can generally serve to test the null of equality of any two unknown density functions f and g. The simple step promoted here consists in fixing one of the two densities to a desired reference density. The test then evaluates whether or not the distance—in an entropy sense—between the still unknown density, $f(x) \equiv \text{p.d.f.}(X)$, and the reference density, $g(x)$, is so small that the former can be judged equal to the latter. In the present context, I shall use two reference densities giving rise to two tests informally summarized by the two H_0's

$$H_0: \quad \text{p.d.f.}(X) \equiv f(x) = g(x, \mu, \sigma) \equiv \tfrac{1}{\sigma}\varphi\left(\tfrac{x-\mu}{\sigma}\right) \qquad \text{for full samples,} \tag{9}$$

$$H_0: \quad \text{p.d.f.}(X) \equiv f(x) = g(x, \mu, \sigma, \xi) \equiv \frac{\frac{1}{\sigma}\varphi(\frac{x-\mu}{\sigma})}{1 - \Phi(\frac{\xi-\mu}{\sigma})} \qquad \text{for deficient samples.} \tag{10}$$

where $\varphi(\cdot)$ and $\Phi(\cdot)$ are the usual notations for the normal p.d.f. and c.d.f., respectively.

Under the null hypothesis and since the normal distribution is fully characterized by its first two moments, p.d.f.(X) and $g(x, \mu, \sigma)$ must have identical parameters. For the full samples case, I set μ and σ in $g(x, \mu, \sigma)$ to be equal to the sample mean and standard deviation of X, respectively. For the deficient samples case, I assume ξ is known. Hence, under the null, μ and σ can be estimated by maximum likelihood. It could also be estimated by the quantile bend estimator—a method proposed in the heights' literature by Wachter and Trussell (1982).[7]

3.3 Nonparametric kernel implementation

The tests introduced above involve one unknown density that remains to be estimated—i.e., $f(x) \equiv \text{p.d.f.}(X)$. For this purpose, I rely on a nonparametric kernel-based estimator (Rosenblatt, 1956, Parzen, 1962). The kernel used is of the Gaussian type. This choice is innocuous (Silverman, 1986) and taken purely for computational convenience.

Unless otherwise specified, the bandwidths are of the plug-in kind computed by the method of Sheather and Jones (1991) based on pilot estimates of derivatives. Despite its merits,[8] this choice might not prove optimal under certain circumstances (Loader, 1999). It is, however, maintained to the detriment of potentially more accurate cross-validation methods. This is for three complementary reasons. First, its risk of over-smoothing is more likely to strike when the smoothing problem is complex. Those cases are rare in anthropometrics where full distributions often suffer from milder deficiencies unlikely to affect smoothing—such as skewness or thin/thick tails. Second, a selected set of simulations was run with cross-validated bandwidths and their results did not differ

[7] Notice, however, that this method has become less popular after being shown inaccurate—Heintel (1996), Komlos (2004).

[8] This is particularly clear when compared to rule-of-thumb references for bandwidth selection.

significantly from the adopted ones. Third, cross-validation methods are reputedly very computationally demanding. Given the number of simulations presented in this paper, it would take too much processing time to provide their equivalent with cross-validated bandwidths. Even with the "fast" plug-in methods, the computations on which this paper is built ran for 62143 hours in a multiple-core cluster, and counting.

All the computations were run in R (R Core Team (2012)) thanks to the "NP" package (Hayfield and Racine, 2008, Racine, 2012), which provides an open platform for nonparametric kernel estimation. Notice, incidentally, that such techniques are either not available in the common statistical software Stata; or, for the univariate kernel density estimation, they are implemented in the most unreliable manner (see StataCorp, 2012, "help kdensity").

Turning now to the known density $g(x)$, it is worth mentioning the following points about the way it enters the estimated metric $\hat{S}_\rho$. I used and evaluated two different approaches: I shall refer to them as the *direct* and the *indirect* approaches. In the direct approach, $g(x)$ is the exact value at $X = x$ of the functions given in (9) and (10). In the indirect approach, $g(x)$ is the nonparametric kernel estimate. This was obtained in the same manner as above but with fewer concerns with its validity because of the known nature of the samples on which it is applied. $\hat{g}(x)$ is estimated from a randomly generated sample whose characteristics are defined in (9) and (10).

Two reasons motivate the use of these alternative approaches. First, notice that, to the extent that $\hat{g}(x) \to g(x)$, the two approaches are asymptotically equivalent. The second reason originates in the difficulty of obtaining critical values for $\hat{S}_\rho$. In the indirect way, I opt for a bootstrap procedure with replacement from the *pooled* empirical distributions (Racine, 2012). Hence the need to generate an "empirical" distribution for the reference distributions defined in (9) and (10). For the direct approach, I calculate 1000 values of $\hat{S}_\rho$ under the null, order them and use, e.g., the 950th value as the 5% critical value.

The use of the test in deficient samples raises a further concern. No combination of kernel type and bandwidth selection method is likely to match the sharp increase of the density at ξ.[9] Hence, there will be a bias in the estimated density $\hat{f}(x)$. However, in the indirect approach, this bias is analogously to be found in the estimated $\hat{g}(x)$. In the direct approach, the bias in $\hat{f}(x)$ will likely increase the $\hat{S}_\rho$ statistic for a sample as much as for the 1000 values of $\hat{S}_\rho$ under the null.

4 Size and power investigations

4.1 Simulation specifics

In this section, I use Monte Carlo simulations to evaluate the size and the power of the proposed tests—the direct $S_{\rho,D}$ and the indirect $S_{\rho,I}$. To better appreciate their performance, I calculate and compare, on strictly identical samples, the results of some popular tests, namely the Jarque-Bera test (Jarque and Bera, 1987, denoted "JB" below), the Shapiro-Wilk test (Shapiro and Wilk, 1965, SW)

[9] I conjecture this would possibly be achieved by a constrained density estimation.

and the d'Agostino test (d'Agostino, 1971, dA). Results for further parametric tests were also calculated and are available upon request. Their performance is in line with those displayed here. For each alternative population and, for each of these latter, each sample size, 1000 samples were drawn and used to gauge the average performance.

Several alternatives are considered. These include three t-distributions ($df = 5, 10, 25$), the uniform distribution and two bi-modal, lightly "contaminated" distributions. They also include a 5% left-truncated distribution for which the problem of truncation is ignored—i.e., the distribution is considered as a full one. Finally, one of the alternatives (Selection I) reflects a process of sample selection. Suppose the underlying population is standard normal. However, observations with higher values have less chance of making it into the sample. Then this would translate into a bias of the estimated parameters with respect to their true value in the population (Heckman, 1979). This process of sample selection can be interpreted in the following way. Changes in the fundamentals of the economy may increase the returns to an individual's height—a proxy for some rewardable feature—in the civilian job market, with respect to its return in the military job market. Therefore, samples of soldiers will exhibit under-representation of tall individuals. In the process studied here, the probability for an observation with normalized value x to be in the sample is $1 - \Phi(x)$.

At this stage, no strict attempt is made to group these alternatives into families based on common criteria—e.g., support, symmetry, skewness, etc. Instead, the set of populations analysed suffices to reveal the potential of these new tests.

The sizes of the samples tested here deserve special attention. They range from 100 to 5000 observations: the list is {100, 200, 300, 400, 500, 1000, 1500, 2000, 3000 and 5000}. This greatly differs from the typically small values chosen in related studies. [10] This choice builds on the following rationale. Firstly, samples of hundreds and even thousands of observations are commonplace in the heights literature (Floud et al., 1990, Nicholas and Steckel, 1991). Also, it does no harm to a test's performance to use it in larger samples—unit power in small samples carries over to larger samples. Moreover, an unsatisfactorily low power can turn into an appropriate power in larger samples. Finally, it is well established that nonparametric, kernel density estimators, such as those used in these news tests, typically require samples that are larger than those providing asymptotic confidence in parametric settings.[11]

4.2 Simulation results

[10] For instance, compare with the sets of the few following studies: {10, 15, 20, 35, 50} in Shapiro et al. (1968), {20, 50, 100} in Pearson et al. (1977), {20} in Arizono and Ohta (1989) or {10, 20, 30, 50} in Noughabi and Arghami (2011).

[11] Notice, however, that such confidence may very well be a deceptive one. Fast convergence ought barely to be an argument if it leads to the wrong parameter. As noted by Robinson (1988), parametric estimators are typically "$\sqrt{n}$-inconsistent".

Table I: Empirical size estimates for $\alpha = 5\%$ tests—Full samples

Population	Test	Sample size									
		100	200	300	400	500	1000	1500	2000	3000	5000
	JB	**.052**	.029	**.050**	**.047**	.036	**.049**	.059	**.046**	.056	**.044**
	SW	.064	.038	.051	**.053**	**.050**	**.049**	.061	.044	**.049**	**.044**
$N(0,1)$	dA	.068	**.039**	.052	.059	.041	.053	.064	.044	.062	.042
	$S_{\rho,D}$	**.052**	**.039**	.054	**.053**	**.050**	.052	**.053**	.030	**.049**	.057
	$S_{\rho,I}$	.007	.006	.007	.005	.004	.009	.014	.020	.018	.015

Notes: – Boldface signals the test of normality with best performance for each pair of alternative population and sample size—unless all tests provide equal results. The Jarque-Bera test (Jarque and Bera, 1987) is denoted "JB", the Shapiro-Wilk test (Shapiro and Wilk, 1965), "SW", and the d'Agostino test (d'Agostino, 1971), "dA". The direct and indirect approaches for my test—$S_{\rho,D}$ and $S_{\rho,I}$—are described in subsection 3.3. This applies to the tables below too.

The results of the simulations described above are given in Tables I, II, III and IV.[12] They give the estimates for the empirical size and power, first for the full samples and then for the deficient ones. I discuss them in turn. Notice that, throughout all the tables, the nominal level of the tests is 5%. Table I shows that the normality test proposed here—calculated in the direct approach—has extremely precise size. The test obtained in the indirect approach seems a bit conservative: it rejects a few more samples than expected at 95%. In order to redress this under-size, the role of the length of the generated twin sample to which the analysed sample is compared should be investigated.

Table II evaluates the power of the test in comparison to the alternatives described above. Overall, the power of the direct test is in line with the power of the parametric tests despite lagging slightly in small samples. Given the sample sizes analysed here, it appears to reach unit power more or less as early as the fast, parametric tests. This result represents a valuable attainment on its own.

Turning to the test for deficient samples, in Table III I consider various values of ξ to evaluate its sizes. Recall that ξ is the truncation point. In the heights literature, it refers to the minimal height requirement imposed by military institutions. Roughly speaking, the results are good. The direct test is undersized for $\xi = -3, -2, -1$ while the indirect test seems very well calibrated. For $\xi = 0$ which is the case when half of a normal distribution is cut off, the tests tend to reject more big samples than expected.

The power investigations for this test, Table IV, include various values of ξ when the population is from the Student type with 5 degrees of freedom. The power of the direct test increases rapidly with the sample size and reaches one in samples of two thousand observations or beyond. Notice that all the tests on the full distribution of that population attain unit power only for samples of one thousand observations or more (see Table II). Therefore, the performance of this test against this particular alternative is good. Unsurprisingly, the power estimate is negatively linked to ξ: the more amputated the distribution, the lower the power.

Two complementary comments are the following. First, as alluded to above, optimal cross-validated bandwidth selection may also be a useful power-enhancing improvement to these tests. Nevertheless, the evidence gathered so far already allows us to paint an overall satisfactory picture based on performance as good as the best parametric competitors and good behaviour in deficient samples. Second, I treat heights as a continuous variable. Readers familiar with the present literature could argue that most of the available samples use rounded data. One way to overcome this issue consists in adding appropriate uniform shocks to the reported rounded heights. Exploratory simulations show that the tests keep their demonstrated properties.

5 Secular trend of British recruits' average height

Floud et al. (1990) offered a path-breaking contribution to the field of anthropo-

[12] Notice that the typical time span for a run of simulations ranged from four to ten days, depending on the number of processes analysed and the availability of the cluster.

Table II: Empirical power estimates for $\alpha = 5\%$ tests—Full samples

Population	Test	Sample size									
		100	200	300	400	500	1000	1500	2000	3000	5000
	JB	**.637**	**.860**	**.949**	**.978**	**.995**	1.00	1.00	1.00	1.00	1.00
	SW	.568	.816	.920	.967	.990	1.00	1.00	1.00	1.00	1.00
$t(5)$	dA	.613	.834	.928	.968	.992	1.00	1.00	1.00	1.00	1.00
	$S_{\rho,D}$	.492	.715	.869	.943	.980	1.00	1.00	1.00	1.00	1.00
	$S_{\rho,I}$	.281	.567	.707	.851	.914	1.00	1.00	1.00	1.00	1.00
	JB	**.275**	**.474**	**.590**	**.644**	**.746**	**.941**	**.990**	**.997**	**1.00**	**1.00**
	SW	.236	.382	.492	.552	.659	.900	.980	.995	.999	**1.00**
$t(10)$	dA	.259	.437	.542	.590	.694	.922	.986	.996	.999	**1.00**
	$S_{\rho,D}$	.189	.279	.358	.418	.513	.787	.947	.975	.996	**1.00**
	$S_{\rho,I}$	.048	.111	.145	.223	.238	.554	.770	.883	.982	.999
	JB	**.105**	**.146**	**.177**	**.222**	**.262**	**.371**	**.477**	**.596**	**.716**	**.897**
	SW	.079	.121	.137	.152	.203	.303	.396	.460	.620	.814
$t(25)$	dA	**.105**	.139	.155	.191	.226	.325	.434	.538	.682	.883
	$S_{\rho,D}$	.064	.091	.087	.099	.128	.177	.257	.292	.425	.646
	$S_{\rho,I}$	.013	.013	.022	.021	.031	.072	.068	.118	.207	.341
	JB	.564	**1.00**	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	SW	.993	**1.00**	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
$U(0,1)$	dA	**.996**	**1.00**	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	$S_{\rho,D}$	.937	**1.00**	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	$S_{\rho,I}$	.794	.999	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	JB	.061	.095	.130	.158	.179	.338	.475	.612	.784	.946
	SW	**.079**	.101	**.140**	**.171**	**.191**	.333	**.476**	.611	.776	.934
Selection I[a]	dA	.075	**.110**	**.140**	.166	.186	**.340**	.472	**.615**	**.787**	**.947**
	$S_{\rho,D}$	.062	.072	.103	.114	.120	.243	.347	.412	.587	.824
	$S_{\rho,I}$	.008	.009	.020	.024	.029	.073	.139	.166	.283	.480
	JB	.092	.371	.640	.872	.967	1.00	1.00	1.00	1.00	1.00
$N(0,1)$	SW	**.315**	**.789**	**.971**	**.997**	**1.00**	1.00	1.00	1.00	1.00	1.00
trunc.	dA	.196	.519	.732	.920	.980	1.00	1.00	1.00	1.00	1.00
5%	$S_{\rho,D}$	.170	.582	.911	.989	**1.00**	1.00	1.00	1.00	1.00	1.00
	$S_{\rho,I}$	.045	.221	.552	.849	.959	1.00	1.00	1.00	1.00	1.00
	JB	**.059**	.060	.063	.072	.087	.118	.143	.168	**.242**	**.375**
	SW	**.059**	.074	.064	**.075**	.088	**.132**	**.147**	**.176**	.215	.325
Mix I[b]	dA	.068	**.080**	**.069**	**.075**	**.092**	.119	.144	.169	.240	.374
	$S_{\rho,D}$	.057	.055	.063	.068	.071	.090	.099	.098	.122	.206
	$S_{\rho,I}$	.015	.034	.047	.029	.033	.031	.037	.042	.054	.138
	JB	.226	.440	.624	.744	.850	**.985**	**1.00**	**1.00**	**1.00**	**1.00**
	SW	**.252**	.480	**.636**	**.775**	**.854**	.984	**1.00**	**1.00**	**1.00**	**1.00**
Mix II[c]	dA	.244	**.452**	.628	.749	.848	.984	**1.00**	**1.00**	**1.00**	**1.00**
	$S_{\rho,D}$	.163	.360	.495	.654	.747	.966	.998	**1.00**	**1.00**	**1.00**
	$S_{\rho,I}$	.064	.116	.224	.274	.344	.489	.632	.721	.754	.812

[a] $f(x) = 2\varphi(x)(1 - \Phi(x))$ where $\varphi(\cdot)$ and $\Phi(\cdot)$ are the p.d.f. and c.d.f. of the standard normal distribution, respectively.

[b] Mix I refers to a population drawn from two normal distributions. With probability p, the observation come from a $N(\mu_1, \sigma^2)$ and the complement in the sample from a $N(\mu_2, \sigma^2)$. For this population, I set $p = .90$, $\mu_1 = 0$, $\sigma^2 = 1$ and $\mu_2 = \mu_1 + 1 \cdot \sigma = 1$.

[c] Same as for Mix I but $\mu_2 = \mu_1 + 2 \cdot \sigma = 2$.

Table III: Empirical size estimates for $\alpha = 5\%$ tests—Deficient samples

Population	ξ	Test	Sample size									
			100	200	300	400	500	1000	1500	2000	3000	5000
$N(0,1)$	-3	$S_{\rho,D}$	.003	.000	.007	.002	.003	.002	.002	.002	.005	.006
		$S_{\rho,I}$	.070	.074	.064	.053	.049	.055	.044	.061	.053	.056
	-2	$S_{\rho,D}$	.000	.000	.001	.003	.001	.003	.005	.011	.003	.011
		$S_{\rho,I}$	.065	.061	.043	.063	.057	.061	.063	.057	.070	.051
	-1	$S_{\rho,D}$	.006	.007	.021	.019	.019	.031	.036	.039	.039	.029
		$S_{\rho,I}$	.065	.051	.044	.054	.036	.044	.034	.053	.058	.048
	0	$S_{\rho,D}$	.005	.007	.010	.015	.026	.062	.107	.156	.264	.356
		$S_{\rho,I}$	.009	.016	.019	.019	.027	.060	.090	.104	.171	.232

Table IV: Empirical power estimates for $\alpha = 5\%$ tests—Deficient samples

Population	ξ	Test	Sample size									
			100	200	300	400	500	1000	1500	2000	3000	5000
$t(5)$	-3	$S_{\rho,D}$	.076	.159	.269	.405	.528	.930	.990	1.00	1.00	1.00
		$S_{\rho,I}$	.170	.292	.454	.597	.712	.956	.998	1.00	1.00	1.00
	-2	$S_{\rho,D}$	.039	.042	.079	.091	.133	.419	.684	.886	.986	.999
		$S_{\rho,I}$	.135	.257	.393	.545	.652	.945	.994	1.00	1.00	1.00
	-1	$S_{\rho,D}$	.038	.078	.151	.211	.254	.610	.852	.949	.996	1.00
		$S_{\rho,I}$	.064	.138	.204	.294	.410	.802	.964	.991	.999	1.00
$\Gamma(9,2)$	1.5	$S_{\rho,D}$	.007	.004	.008	.007	.015	.056	.185	.395	.776	.994
		$S_{\rho,I}$	.043	.131	.201	.346	.485	.901	.987	.999	1.00	1.00
	3	$S_{\rho,D}$	.006	.008	.011	.024	.019	.053	.087	.145	.211	.419
		$S_{\rho,I}$	.006	.006	.007	.013	.020	.038	.112	.207	.384	.612
$U(-3,3)$	$-2\sqrt{3}$	$S_{\rho,D}$	.223	.735	.972	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		$S_{\rho,I}$	.585	.979	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	$-\sqrt{3}$	$S_{\rho,D}$	.640	.993	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		$S_{\rho,I}$	.809	.999	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

metric history. In a joint project (Fogel et al., 1983) they collected and examined a formidable data set on the heights—and further covariates such as age—of recruits to both the British Army and the Royal Marines between 1760 and 1889. A direct implication of their work has been a new assessment of the secular change in the British stature. They find a significant positive impact of the Industrial Revolution on average heights (although with some regression in the mid-XIX century); and they find convergence between social classes.

This optimistic conclusion, however, has been fiercely challenged by numerous scholars who find instead a negative impact of the Industrial Revolution on workers' living standards (Mokyr, 1988). Using an anthropometric approach with the same or different data sets, Komlos (1993, 1998), Komlos and Kuechenhoff (2012) and Nicholas and Steckel (1991) both find a negative effect of the Industrial Revolution on heights.

In this section I re-evaluate the validity of the Floud et al. (1990) results using the tests developed above and the dataset they have collected and kindly deposited at the UK data archive (Floud, 1986). I shall proceed in the following way. Floud et al. (1990) derive height estimates for quinquennial samples using a maximum-likelihood estimator given by

$$\mathcal{L}(x_i, \mu, \sigma, \xi) = \prod_{i=1}^{n} \frac{\frac{1}{\sigma}\varphi\left(\frac{x_i-\mu}{\sigma}\right)}{1-\Phi\left(\frac{\xi-\mu}{\sigma}\right)}.$$

Clearly, if a quinquennial sample is not normally distributed, then the estimated $\hat{\mu}$ can be misleading. It violates the assumption that it is randomly drawn from a normally distributed population. The tests presented above allow me to check whether or not normality can be rejected in the sample. If the normality assumption is rejected then I suggest that the estimated mean of the sample is unreliable and should be discarded. Only valid samples should be used to infer trends in heights.

5.1 Defining the samples

Various ways exist to draw (sub-)samples from any historical dataset under examination. Observations are generally grouped into birth cohorts in the height literature. However, different authors have chosen different criteria for a) the time span of the samples; and b) the age at recruitment of the individuals to be included. For instance, Komlos (1993) uses decades for criterion a) while Floud et al. (1990) chooses five-year periods. If the samples are large enough then estimates can be obtained for every age at recruitment, to answer criterion b). But it is common to group individuals according to their age at recruitment to get larger samples. Floud et al. (1990) provide estimates for each individual age between 18 and 23 but also group together individuals aged 18-19, 21-23 and 24-29 years (Floud et al., 1990, pp. 136-138). In the present study, and for comparison purposes, I adopt the Floud et al. (1990) five-year birth cohorts and the following age groups: 18-19, 20-22 and 23-24. Comparison with the Floud et al. (1990) estimates is made possible by averaging the corresponding yearly values given by their Table 4.1, pp. 140-149. Notice, as well, that the choice of the

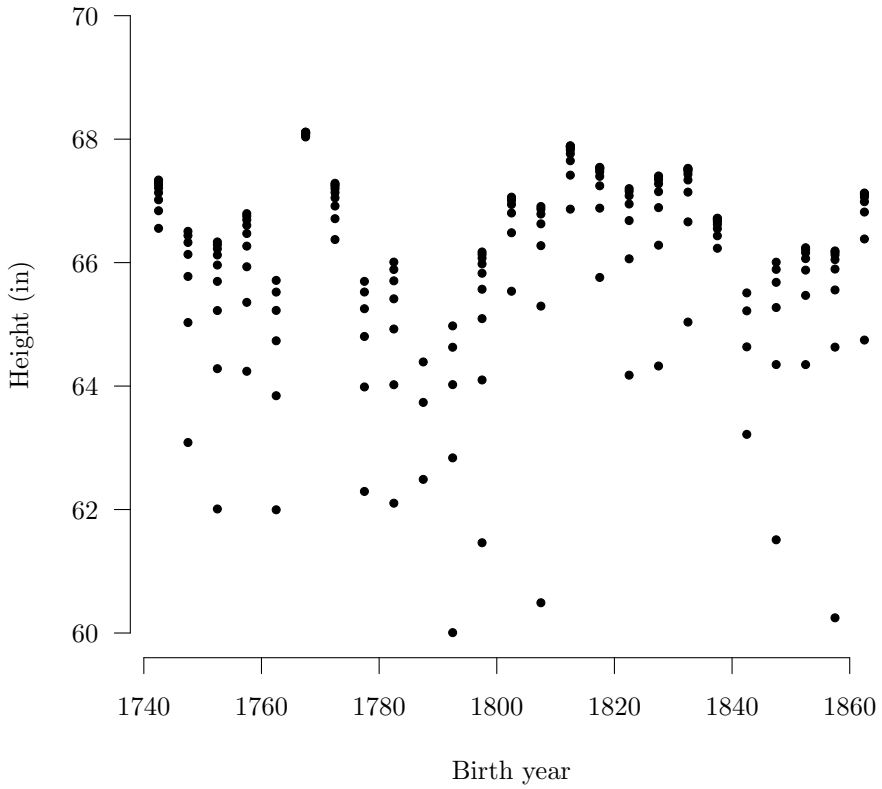

Figure 1: Estimated average heights for different truncation points. Age group 18-19 at recruitment.

18-19 and 23-24 age groups obeys another criterion: they are a quinquennium apart. Although their age at recruitment differs they all have the same birth year. Thus any difference in average height between these two groups could not be linked to different social and economic conditions during the growth period of the individuals. Instead, it would indicate that individuals did not reach their adult height at the age of 18 or 19.

5.2 Choice of truncation points

The minimum height requirement imposed by many armies would ideally translate into a clear fall (to zero) in the left part of the height distribution. However, this ideal case is undermined by various factors such as laxity in regulation enforcement or changes over time of the minimum height requirements. Overall, the resulting height distributions show truncation on the left tail but leave some observations to the left of the truncation point—a problem known as shortfall (Wachter and Trussell, 1982). This problem has been addressed by Wachter and Trussell (1982) who proposed the quantile bend estimator.

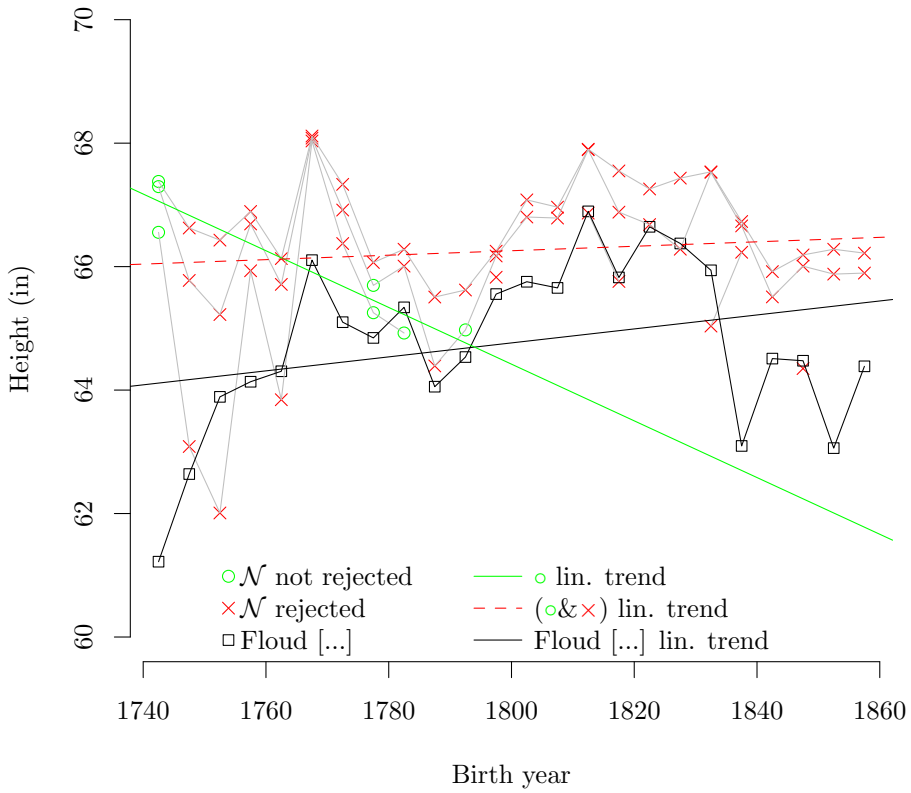

Figure 2: Mean height of 18- to 19-year-old recruits.

By contrast, the reduced-sample/truncated maximum likelihood estimator makes no use of the observations to the left of the truncation point. This, in turn, adds considerable importance to an appropriate choice of the truncation point in the sample. To illustrate this idea, I ran the following exercises. For an initial likely truncation point and nine others to its left—height steps of 1/4 inch and one assuming no truncation of the sample—I calculated the mean of the truncated samples. A representative result is given by Figure 1. It clearly shows that the mean of the sample can fluctuate by a large amount depending the truncation point chosen. For example, in 1855-60 we might infer an average height of anything between 60 and 66 inches depending on the truncation point chosen.

Given this sensitivity of the sample mean to the choice of the truncation point, it is unfortunate that Floud et al. (1990) do not provide explicit information about their truncation points. Their discussion of the problem, pp. 132-133, reveals, however, that it was a very complex task to assign a minimum requirement for each observation of their dataset; no clear rule seems to have been followed.

In the present study I rely on the visualisation of the histograms in order to

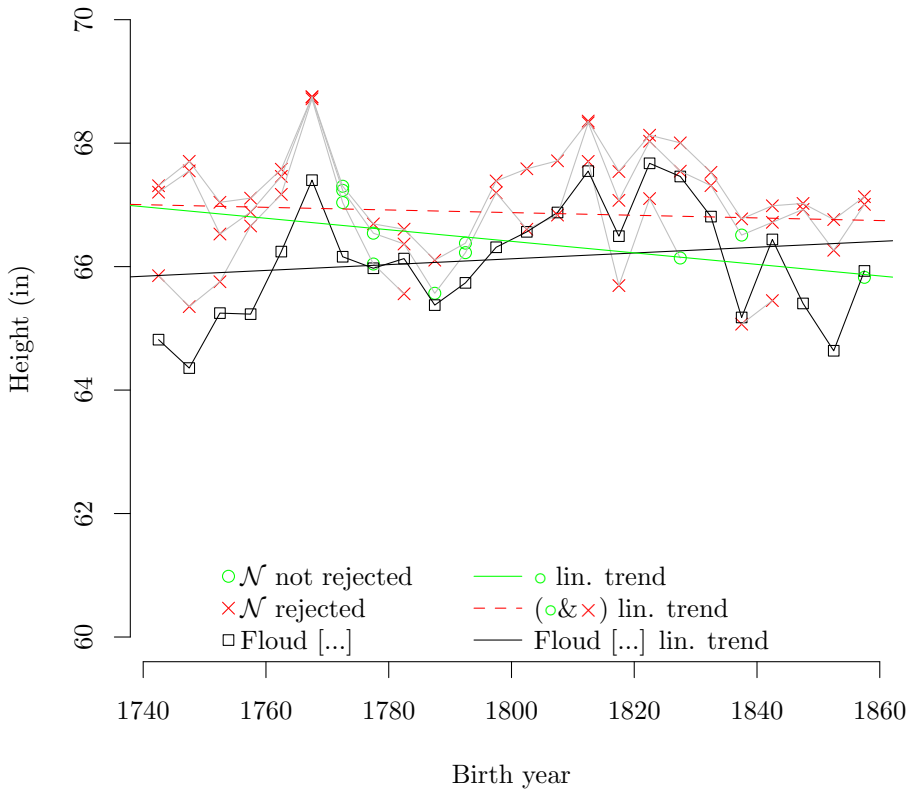

Figure 3: Mean height of 20- to 22-year-old recruits.

assess three likely truncation points. These are given in Appendix A along with the corresponding histograms. A few important points should be noted. Since the real values of the truncation point used during the recruitment process are not known, one has no other choice than try to deduce them from the actual samples. I chose therefore the most transparent way of doing so.[13] Notice, also, that several alternatives were estimated, of the kind illustrated by Figure 1, but no significant differences were found in the results obtained.

5.3 New estimates of recruits' average height

Samples from the Floud et al. (1990)'s dataset were drawn following the criteria given above—five-year birth cohorts and three age-at-recruitment groups—using the truncation points given in Appendix A. For each sample I calculate the mean using the reduced-sample/truncated maximum likelihood estimator. I also use the test proposed above—$S_{\rho,I}$—in order to evaluate whether or not the normality assumption can be rejected in the given sample. Importantly, all the

[13]Notice, incidentally, that Floud et al. (1990) do not provide their estimates of the truncation points.

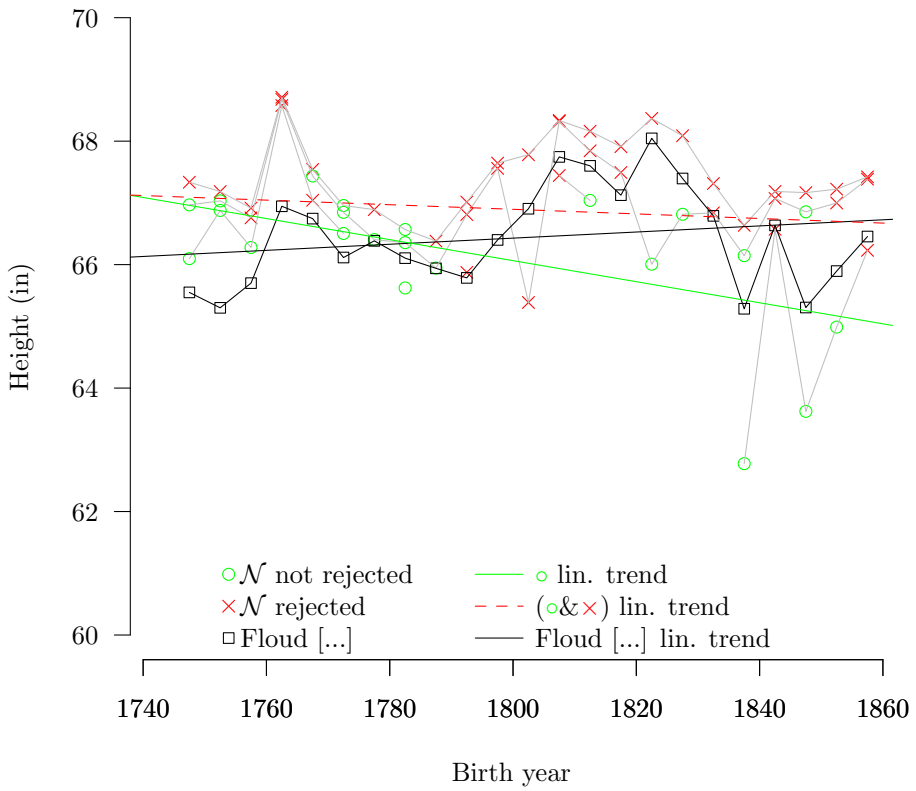

Figure 4: Mean height of 23- to 24-year-old recruits.

samples analysed here are drawn from the infantry—excluding marines.

Results are given in Figures 2, 3 and 4. They show my estimates for the recruits' average height between 1740 and 1860 and compare them with the Floud et al. (1990) estimates. For each year, I give three values, corresponding to the three truncation points identified in Appendix A. Estimates marked with a $\times$ indicated that normality is rejected in the sample—for the given year and truncation point. With a $\circ$ I noted the estimated average height from a sample where normality is not rejected by the test. Floud et al. (1990) estimates are marked with a $\square$.

Lines between points have an indicative purpose only. They connect—over time—estimates for the smallest, middle and highest truncation point of each year. In other words, they draw three possible evolutions of average height to be compared with the line provided by Floud et al. (1990).

Three linear trends are also provided for a) values derived from samples where normality is not rejected; b) values for all my estimates; and c) values of Floud et al. (1990). The purpose here is not to best fit the data points. Instead, they are meant to give a general idea about the secular trend of soldiers average height. The following points are worth noting.

For all age groups, Floud et al. (1990) results show an upward trend of recruits' height over the period. This serves as a benchmark for their main conclusion, despite the fact that they use further datasets—i.e., more data points—and a smoothing technique to evaluate the secular trend.[14]

Secondly, the trends drawn using all of my estimates are qualitatively different from those of Floud et al. (1990). The slope—the red dashed line—for the 18-19 age group is positive like the one in Floud et al. (1990) but it's at odds with this latter in the remaining two groups—though both appear to have a slope close zero. One could test whether these differences are statistically different or not. Also, further structure in the estimated trend would certainly reveal further similarities or disparities in the series. I do not proceed with this comparison since estimates marked with $\times$ are not considered here as reliable.

Third, and most importantly, the normality assumption is rejected in a vast majority of the samples. This implies that the truncated maximum-likelihood estimates generated by Floud et al. (1990) are potentially misleading. In turn, this result casts doubt on any view about the secular trend based on these estimates.

A few differences across age groups are also present. Average height increases with age, suggesting that individuals do not reach their adult height before the age of 20 or more. This point is relevant for studies using younger people (Komlos, 1993, e.g.,). Indeed, the assumption—valid or not—that adult population height is normally distributed does not necessarily imply normality at all ages during the growth period. As far as the results presented here can establish, the assumption of normality in samples of younger people is more easily rejected.

Finally, if one were to restrict the analysis to samples where normality is not rejected—estimates marked with a $\circ$—then the view about the secular trend is even more dramatically different from the one in Floud et al. (1990). The trend, in that case, is given by the green lines. Since these lines build on the observations that are not rejected by the test proposed here, I argue that a downward secular trend in the average recruits' height cannot be rejected and, indeed, is more likely.

The implications of these results extend to the main debate where anthropometric indicators—such as Floud et al. (1990)'s—have brought new insights. Claiming that average heights increased during the late eighteenth and early nineteenth century, Floud et al. (1990) side with the "optimists" to assert the beneficial effects on the working classes of the early stages of the Industrial Revolution. Based on the tests presented here, I argue that Floud et al. (1990)'s results are misleading and of low support for their side of the controversial issue. Put differently, the present results tend to support the "pessimistic" side defended, for instance, by Komlos (1993) or Komlos and Kuechenhoff (2012).

[14]Should one look in more detail at this upward linear trend, one would notice the following. Floud et al. (1990) show a decline in heights after around 1820 but an increase during the key period of the Industrial Revolution.

6 Conclusion

The normality assumption is too pervasive in the anthropometrics literature to be left unchecked in all available samples. In the current practice, however, this fails to apply with full rigour since "statistical tests of normality have not been devised for distributions with height requirements" (Komlos, 1994).

This paper adds two useful tools to the researchers' arsenal of tests aimed at detecting departures from normality. In contrast to the existing parametric tests, the tests proposed here are consistent tests building on a metric entropy based on nonparametrically estimated densities. Importantly, their performance is quite remarkable in simulated data.

The first test applies to full distributions and is shown to have a performance in line with the performance of its parametric counterparts. The second test is alone in its class as it is the first to apply to truncated samples that are commonplace in the field. Size and power investigations show again reasonably good behaviour of the test.

The classic data set of Floud et al. (1990) is re-analysed in the light of these new tests. It is shown that the normal distributional prior adopted by these authors, and the current literature is an inappropriate description of the recruits' height distribution in most cases. This is particularily true for the youngest individuals. The consequence of these tests is quite dramatic. The upward secular trend drawn by the Floud et al. (1990) estimates turns—if one restricts calculations to validly inferred estimated average heights—into a downward secular trend also previously obtained by other scholars like Komlos (1993).

References

A'Hearn, B. (2004). A restricted maximum likelihood estimator for truncated height samples. *Economics & Human Biology*, *2*, 5–19.

Arizono, I., and Ohta, H. (1989). A test for normality based on Kullback-Leibler information. *The American Statistician*, *43*, 20–22.

Bierens, H. J. (1982). Consistent model specification tests. *Journal of Econometrics*, *20*, 105–134.

Bodenhorn, H., Guinnane, T., and Mroz, T. (2012). Sample-selection bias in the historical heights literature. Unpublished manuscript presented at the Cliometrics Conference, Tucson, USA, May 2012.

Cameron, N. (2003). Physical growth in a transitional economy: the aftermath of South African apartheid. *Economics & Human Biology*, *1*, 29–42.

d'Agostino, R. B. (1971). An omnibus test of normality for moderate and large size samples. *Biometrika*, *58*, 341–348.

Floud, R. (1986). *Long-term Changes in Nutrition, Welfare and Productivity in Britain; Physical and Socio-economic Characteristics of Recruits to the Army*

and Royal Marines, 1760-1879. Colchester, Essex: UK Data Archive. SN: 2131.

Floud, R., Wachter, K., and Gregory, A. (1990). *Height, health and history: nutritional status in the United Kingdom, 1750-1980*. Cambridge University Press.

Floud, R., Wachter, K. W., and Gregory, A. (1993). Measuring historical heights—short cuts or the long way round: a reply to Komlos. *The Economic History Review*, *46*, 145–154.

Fogel, R. W., Engerman, S. L., Floud, R., Friedman, G., Margo, R. A., Sokoloff, K., Steckel, R. H., Trussell, T. J., Villaflor, G., and Wachter, K. W. (1983). Secular changes in American and British stature and nutrition. *The Journal of Interdisciplinary History*, *14*, 445–481.

Galton, F. (1875). Notes on the Marlborough school statistics. *The Journal of the Anthropological Institute of Great Britain and Ireland*, *4*, 130–135.

Granger, C., Maasoumi, E., and Racine, J. (2004). A dependence metric for possibly nonlinear processes. *Journal of Time Series Analysis*, *25*, 649–669.

Havrda, J., and Charvát, F. (1967). Quantification method of classification processes. concept of structural α-entropy. *Kybernetika*, *3*, 30–35.

Hayfield, T., and Racine, J. S. (2008). Nonparametric econometrics: The np package. *Journal of Statistical Software*, *27*, 1–32.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, *47*, 153–161.

Heintel, M. (1996). Historical height samples with shortfall: A computational approach. *History and Computing*, *8*, 24–37.

Jacobs, J., Katzur, T., and Tassenaar, V. (2008). On estimators for truncated height samples. *Economics & Human Biology*, *6*, 43–56.

Jarque, C. M., and Bera, A. K. (1987). A test for normality of observations and regression residuals. *International Statistical Review*, *55*, 163–172.

Komlos, J. (1993). The secular trend in the biological standard of living in the United Kingdom, 1730-1860. *The Economic History Review*, *46*, 115–144.

Komlos, J. (1994). On the significance of anthropometric history. *Stature, Living Standards, and Economic Development*, (pp. 210–220).

Komlos, J. (1998). Shrinking in a growing economy? The mystery of physical stature during the Industrial Revolution. *The Journal of Economic History*, *58*, 779–802.

Komlos, J. (2004). How to (and how not to) analyze deficient height samples. *Historical Methods*, *37*, 160–173.

Komlos, J., and Kuechenhoff, H. (2012). The diminution of the physical stature of the English male population in the eighteenth century. *Cliometrica*, *6*, 45–62.

Komlos, J. H., and Kim, J. H. (1990). Estimating trends in historical heights. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, *23*, 116–120.

Li, Q., and Racine, J. S. (2007). *Nonparametric Econometrics: Theory and Practice*. Princeton University Press.

Lindert, P. H., and Williamson, J. G. (1983). English workers' living standards during the Industrial Revolution: A new look. *The Economic History Review*, *36*, 1–25.

Loader, C. R. (1999). Bandwidth selection: Classical or plug-in? *The Annals of Statistics*, *27*, 415–438.

Maasoumi, E. (1993). A compendium to information theory in economics and econometrics. *Econometric reviews*, *12*, 137–181.

Mokyr, J. (1988). Is there still life in the pessimist case? Consumption during the Industrial Revolution, 1790-1850. *The Journal of Economic History*, *48*, 69–92.

Mokyr, J., and Ó Gráda, C. (1996). Height and health in the United Kingdom 1815–1860: evidence from the East India company army. *Explorations in Economic History*, *33*, 141–168.

Nicholas, S., and Steckel, R. H. (1991). Heights and living standards of English workers during the early years of industrialization, 1770-1815. *Journal of Economic History*, *51*, 937–957.

Noughabi, H. A., and Arghami, N. R. (2011). Monte Carlo comparison of seven normality tests. *Journal of Statistical Computation and Simulation*, *81*, 965–972.

Parzen, E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, *33*, 1065–1076.

Pearson, E. S., D'Agostino, R. B., and Bowman, K. O. (1977). Tests for departure from normality: Comparison of powers. *Biometrika*, *64*, 231–246.

Pearson, K. (1895). Contributions to the mathematical theory of evolution. II. Skew variation in homogeneous material. *Philosophical Transactions of the Royal Society of London. A*, *186*, 343–414.

R Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Racine, J. S. (2012). *Entropy-Based Inference using R and the* `NP` *Package: A Primer*. Unpublished Manuscript McMaster University.

Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica*, *56*, 931–954.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, *27*, 832–837.

Shapiro, S. S., and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*, 591–611.

Shapiro, S. S., Wilk, M. B., and Chen, H. J. (1968). A comparative study of various tests for normality. *Journal of the American Statistical Association*, *63*, 1343–1372.

Sheather, S. J., and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society B*, *53*, 683–690.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis* volume 26. Chapman & Hall/CRC.

StataCorp (2012). *Stata 12 Base Reference Manual*. College Station, TX: Stata Press.

Steckel, R. H. (1995). Stature and the standard of living. *Journal of Economic Literature*, *33*, 1903–1940.

Tanner, J. M., Whitehouse, R., and Takaishi, M. (1966). Standards from birth to maturity for height, weight, height velocity, and weight velocity: British children, 1965. *Archives of Disease in Childhood*, *41*, 454–471, 613–635.

Ullah, A. (1996). Entropy, divergence and distance measures with econometric applications. *Journal of Statistical Planning and Inference*, *49*, 137–162.

Wachter, K. W., and Trussell, J. (1982). Estimating historical heights. *Journal of the American Statistical Association*, *77*, 279–293.

Yazici, B., and Yolacan, S. (2007). A comparison of various tests of normality. *Journal of Statistical Computation and Simulation*, *77*, 175–183.

Appendix A

The following figures give the height distribution for different birth cohorts. All distributions display the values in the 60-73 in range. For each birth cohort, three age groups for the age of recruitment are defined: 18-19, 20-22 or 23-24 (from the left to the right. For each age group, three truncations points are visually identified.

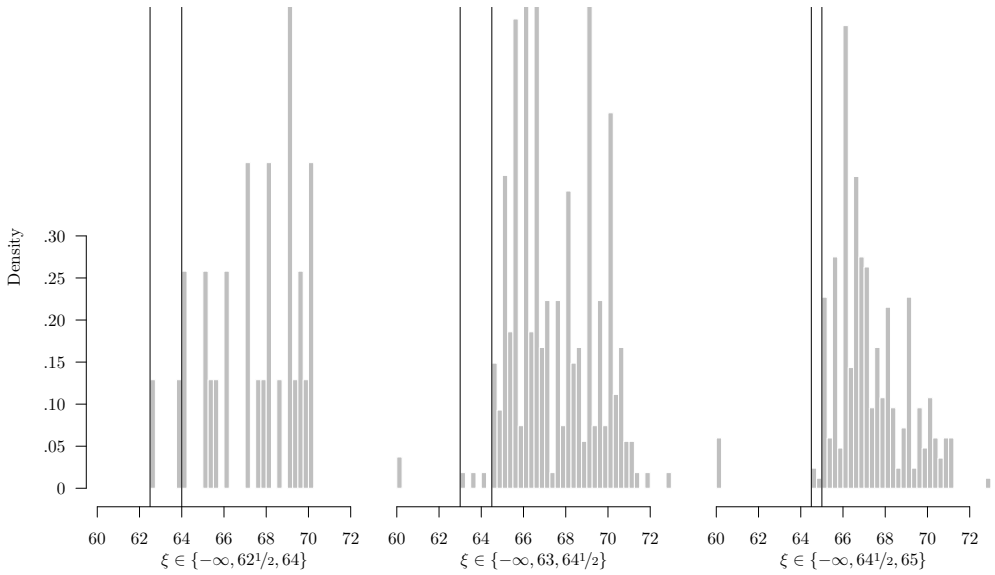

Figure 5: Birth cohort 1740-44.

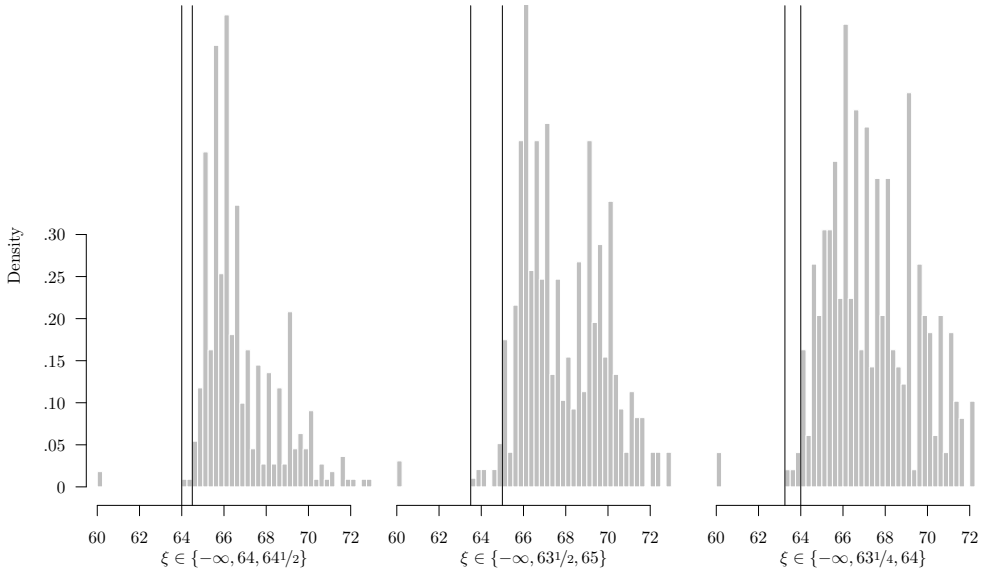

Figure 6: Birth cohort 1745-49.

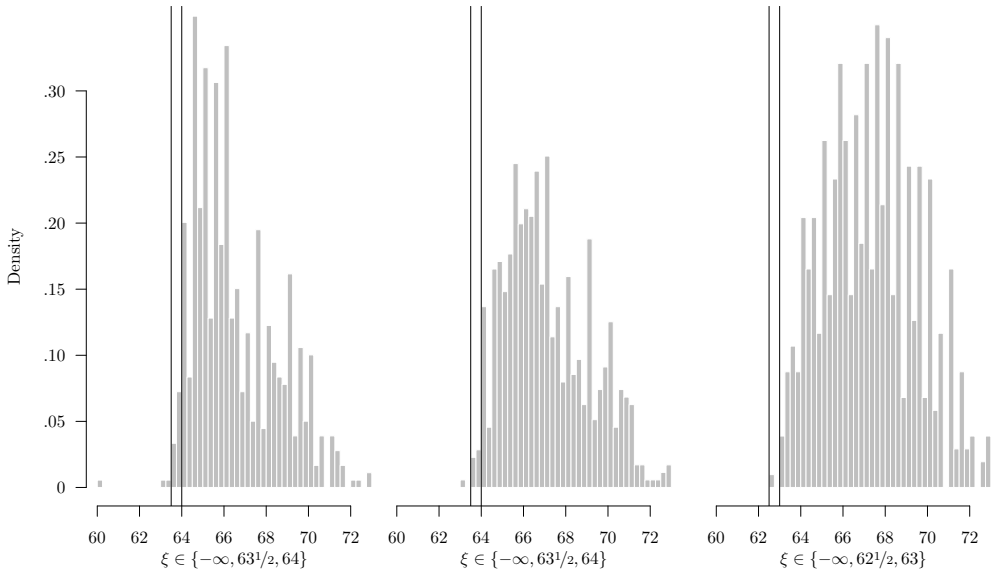

Figure 7: Birth cohort 1750-54.

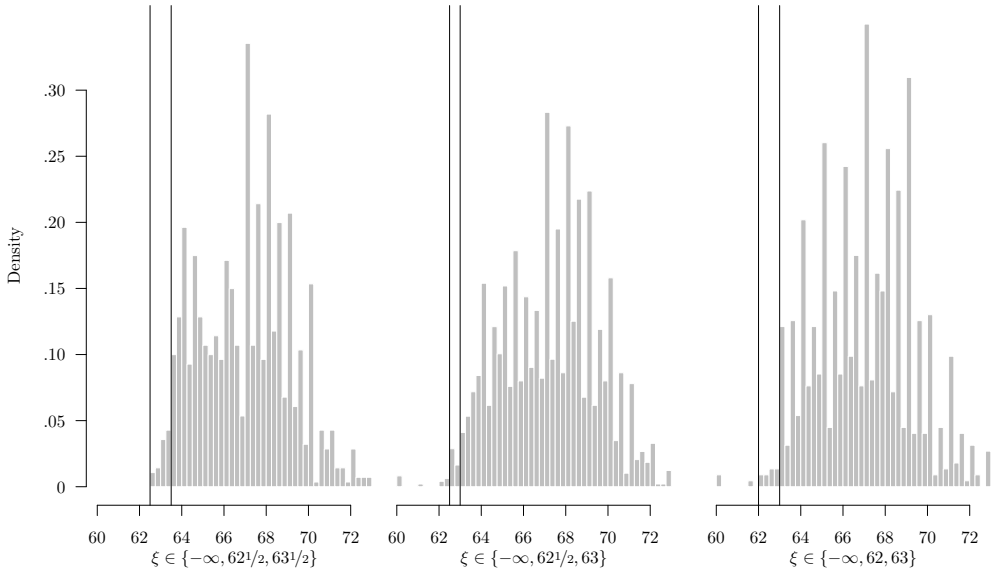

Figure 8: Birth cohort 1755-59.

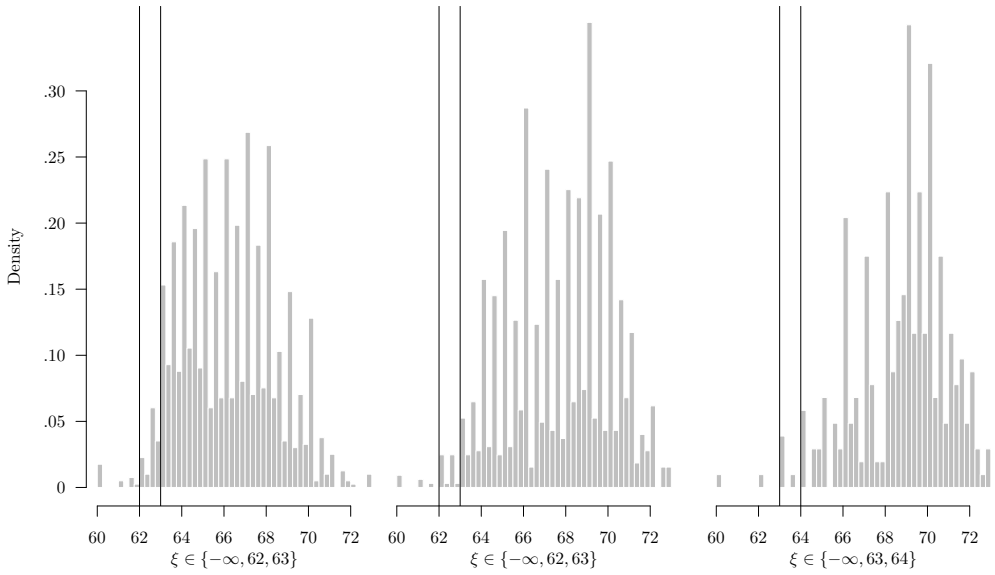

Figure 9: Birth cohort 1760-64.

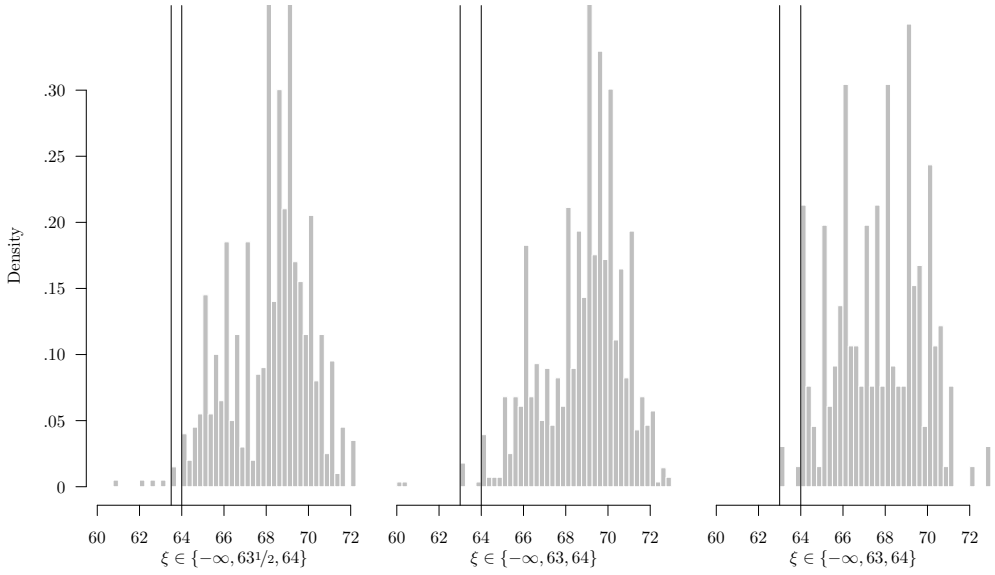

Figure 10: Birth cohort 1765-69.

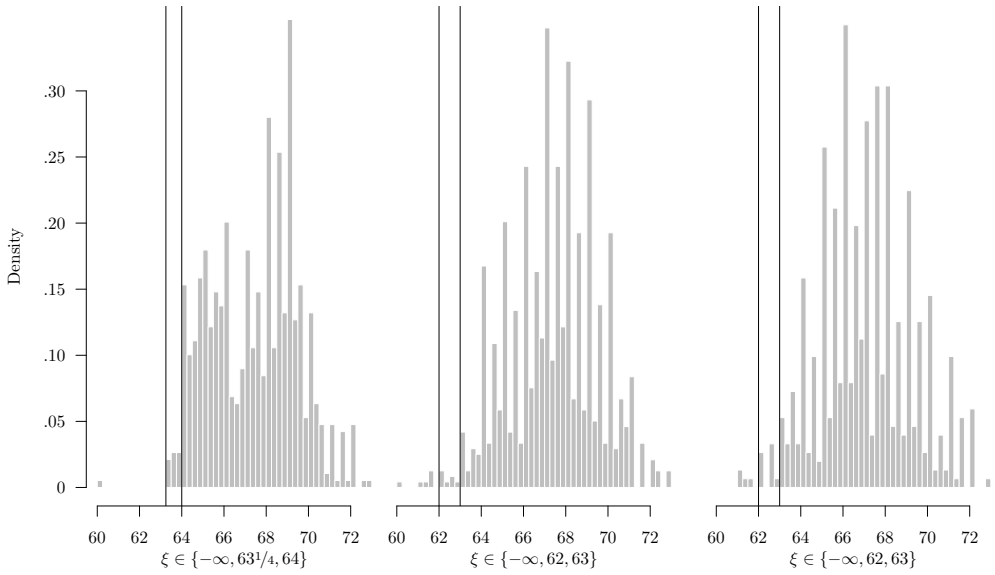

Figure 11: Birth cohort 1770-74.

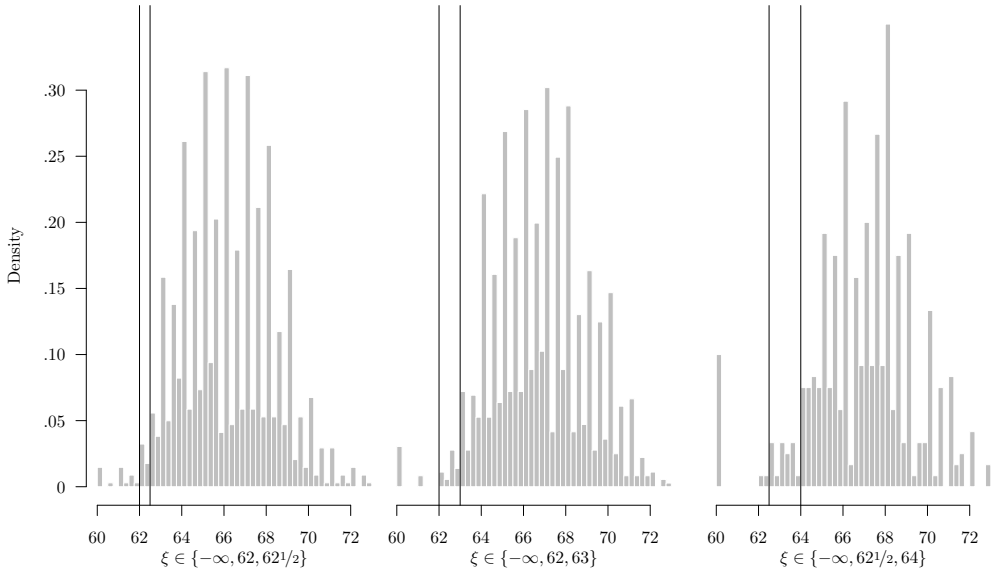

Figure 12: Birth cohort 1775-79.

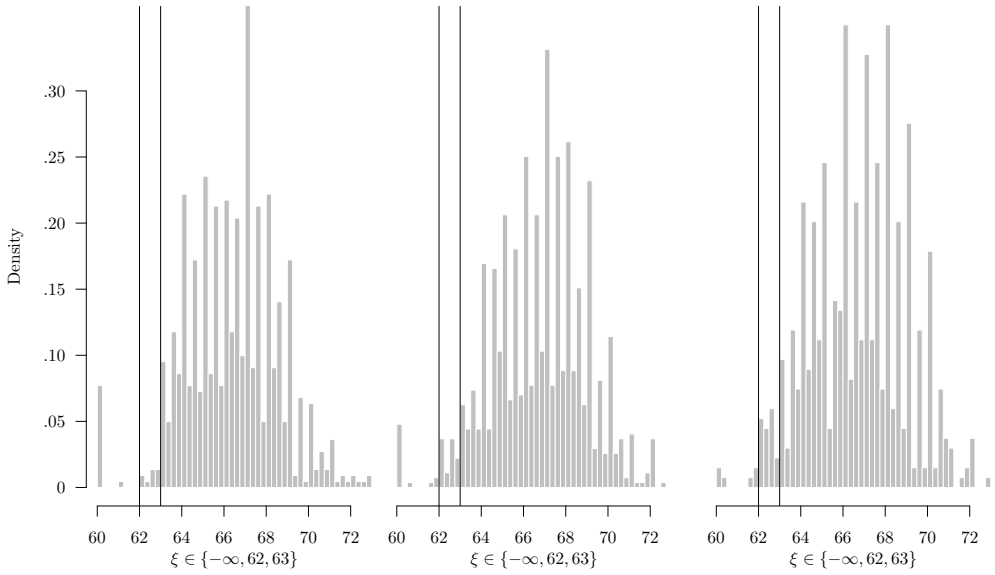

Figure 13: Birth cohort 1780-84.

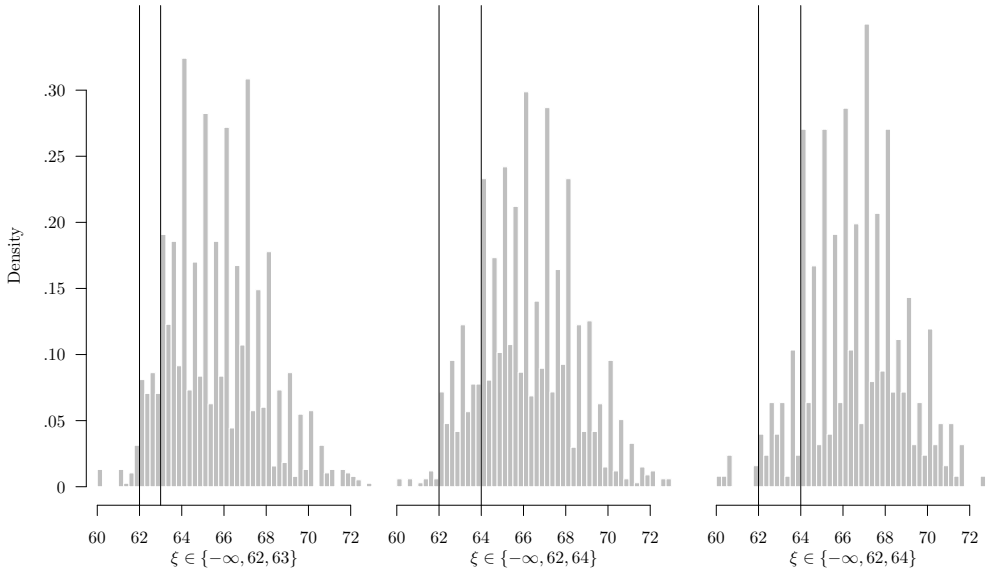

Figure 14: Birth cohort 1785-89.

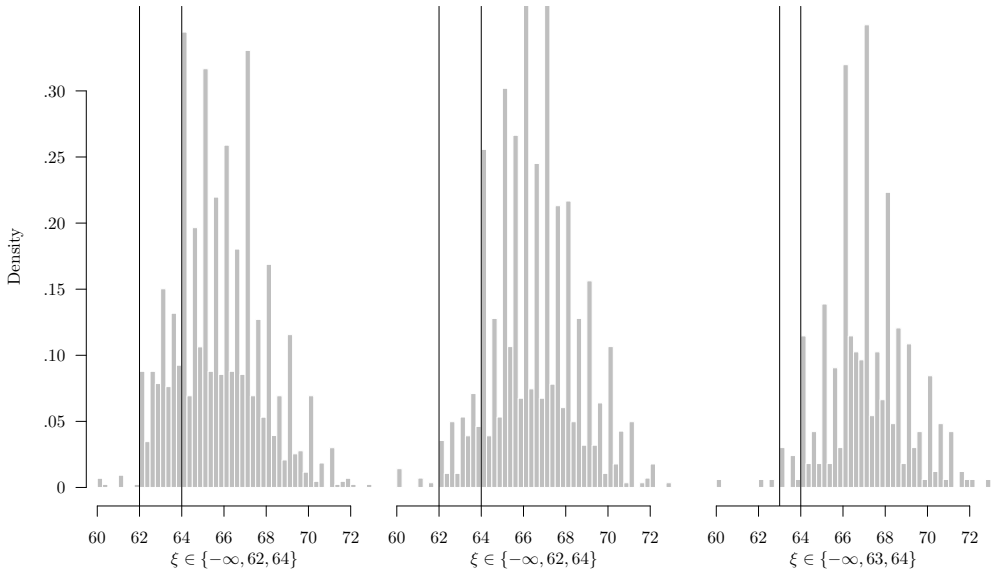

Figure 15: Birth cohort 1790-94.

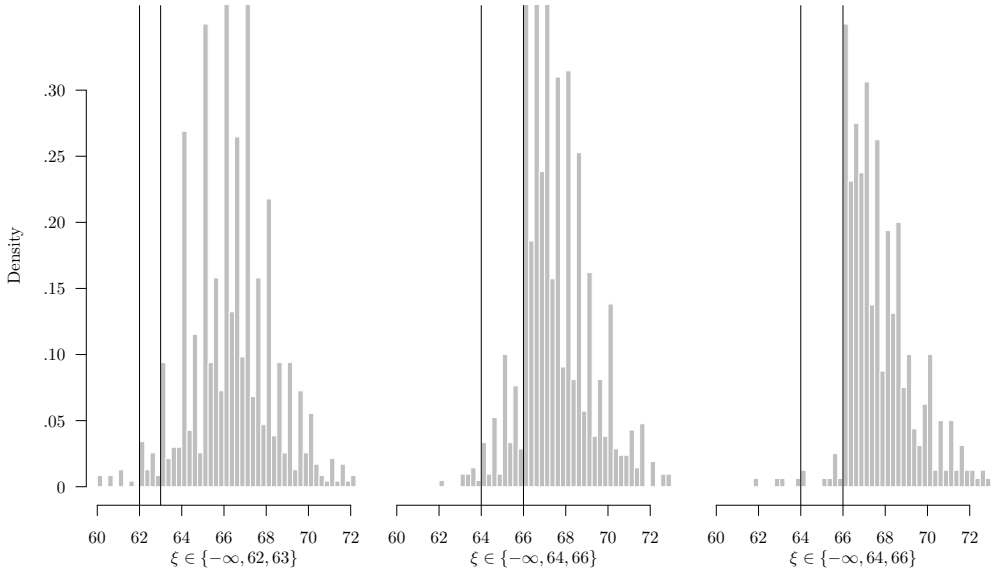

Figure 16: Birth cohort 1795-99.

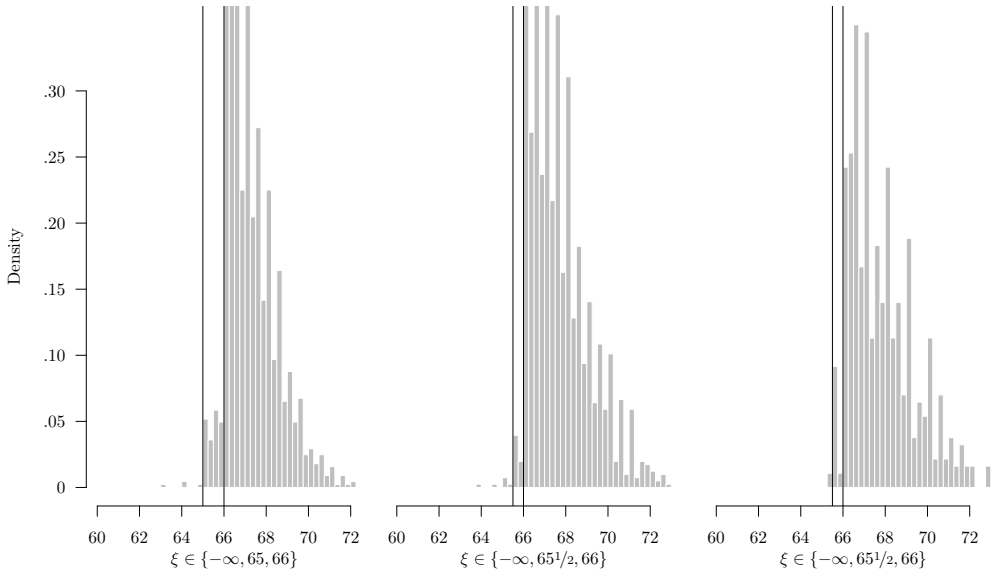

Figure 17: Birth cohort 1800-04.

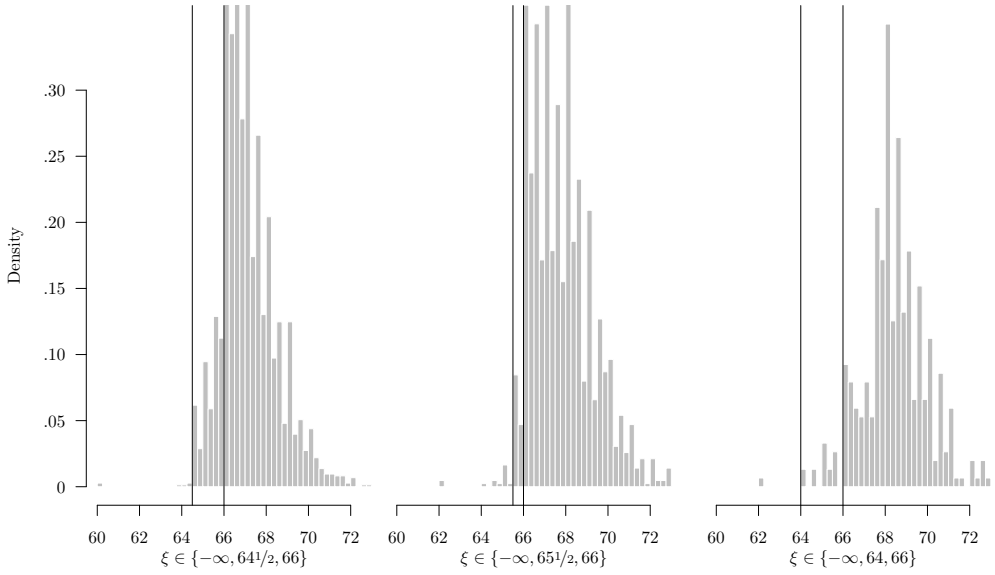

Figure 18: Birth cohort 1805-09.

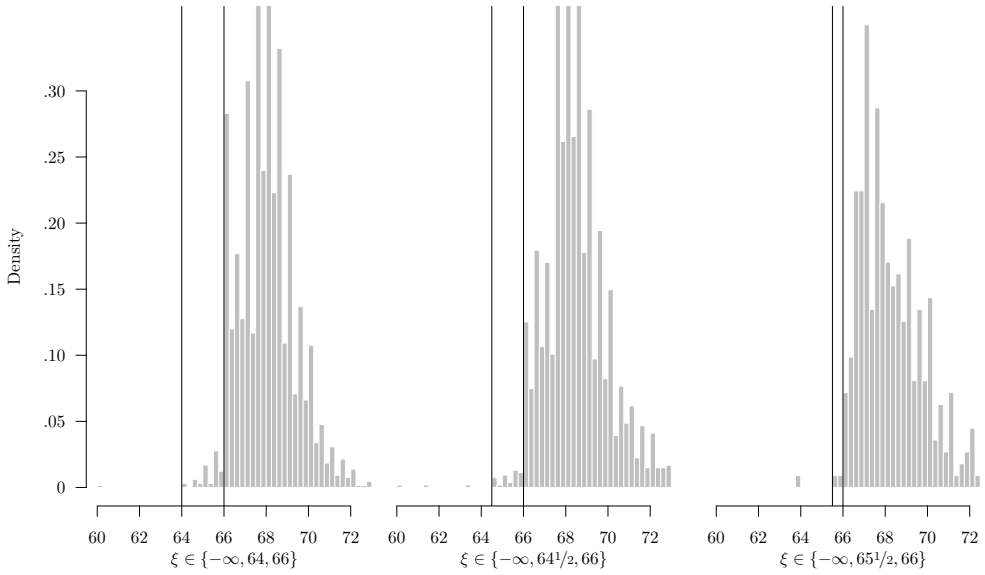

Figure 19: Birth cohort 1810-14.

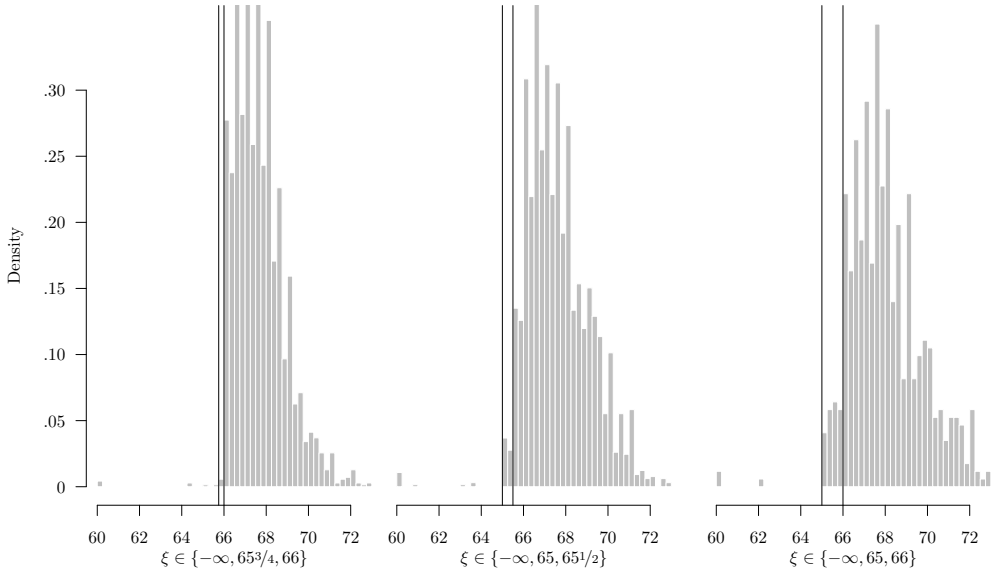

Figure 20: Birth cohort 1815-19.

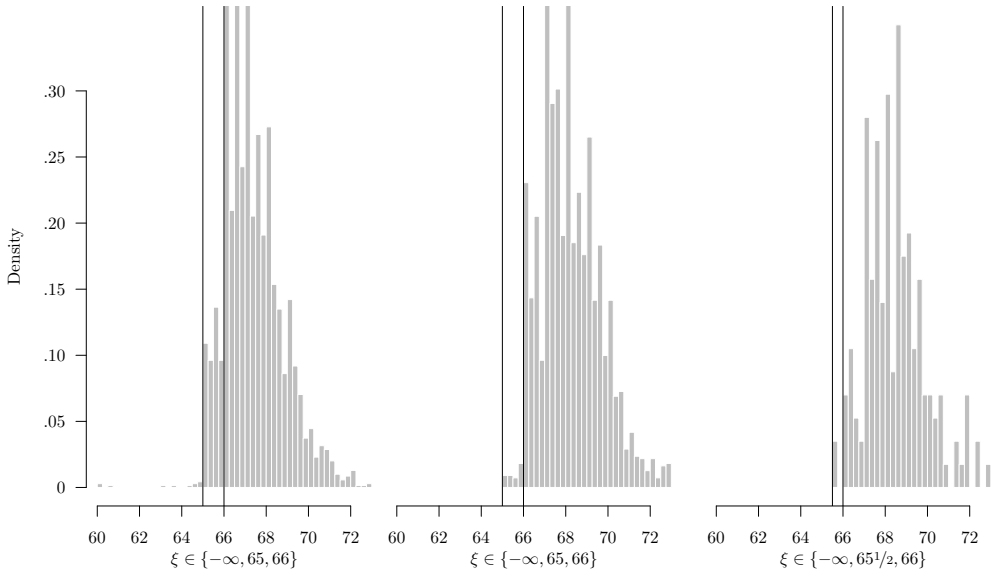

Figure 21: Birth cohort 1820-24.

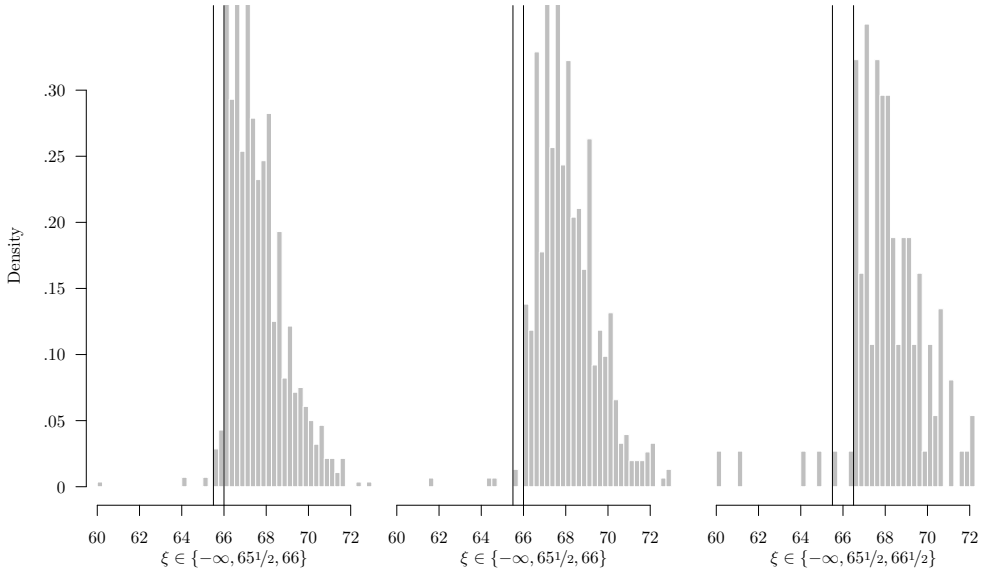

Figure 22: Birth cohort 1825-29.

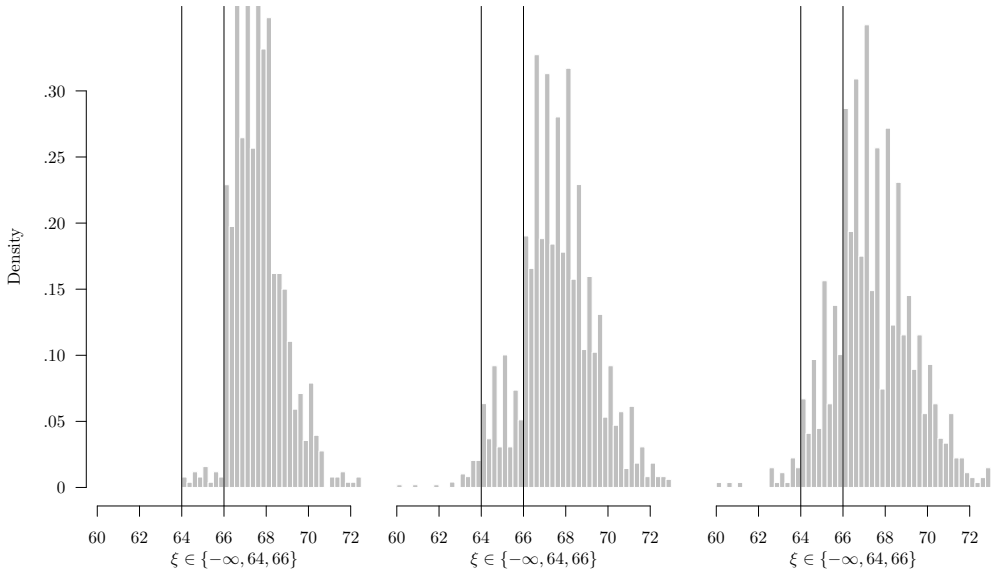

Figure 23: Birth cohort 1830-34.

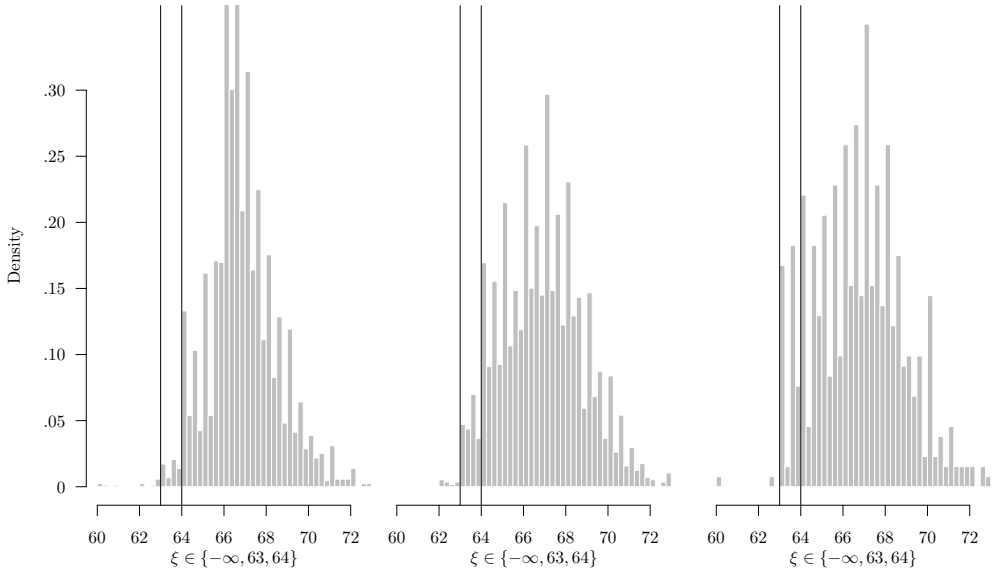

Figure 24: Birth cohort 1835-39.

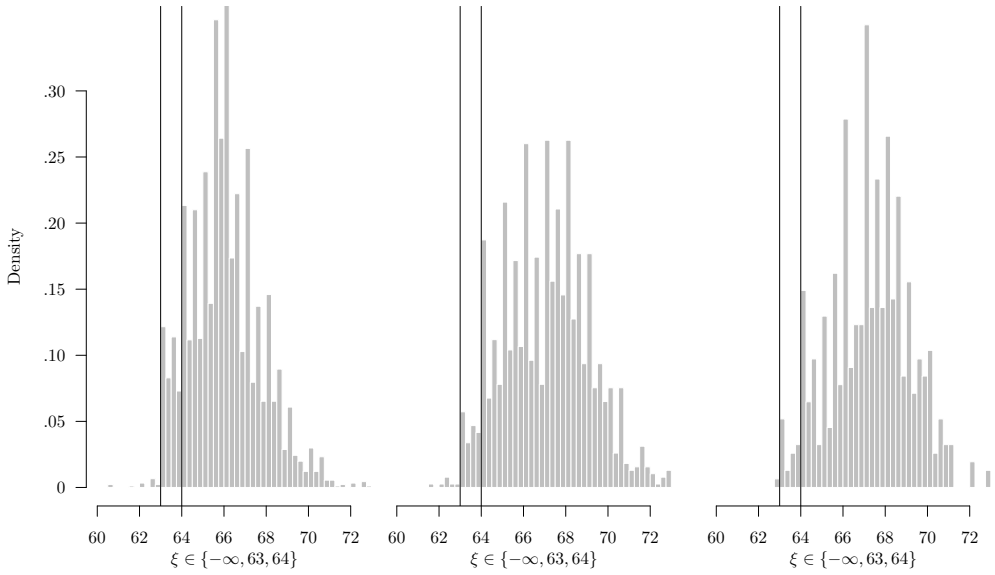

Figure 25: Birth cohort 1840-44.

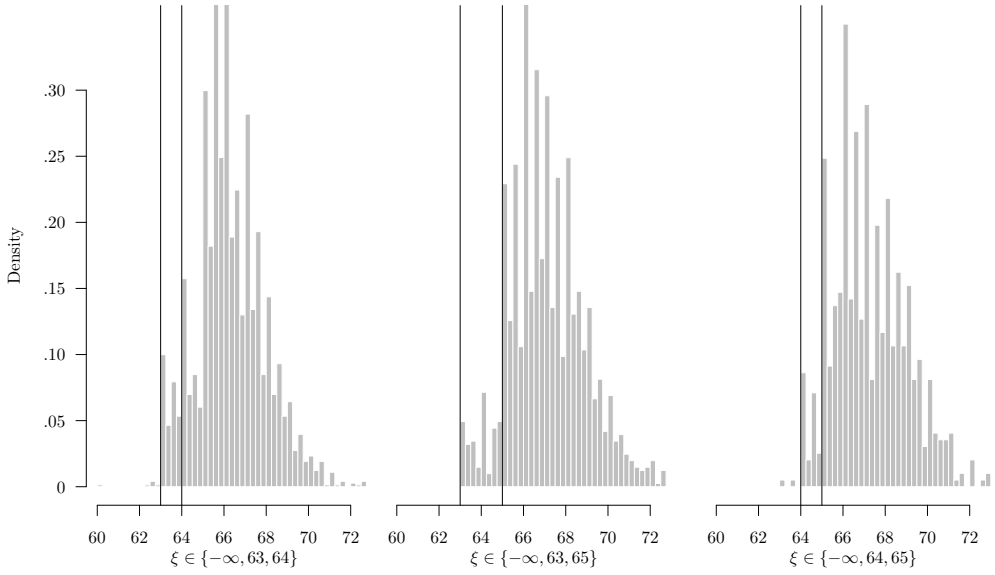

Figure 26: Birth cohort 1845-49.

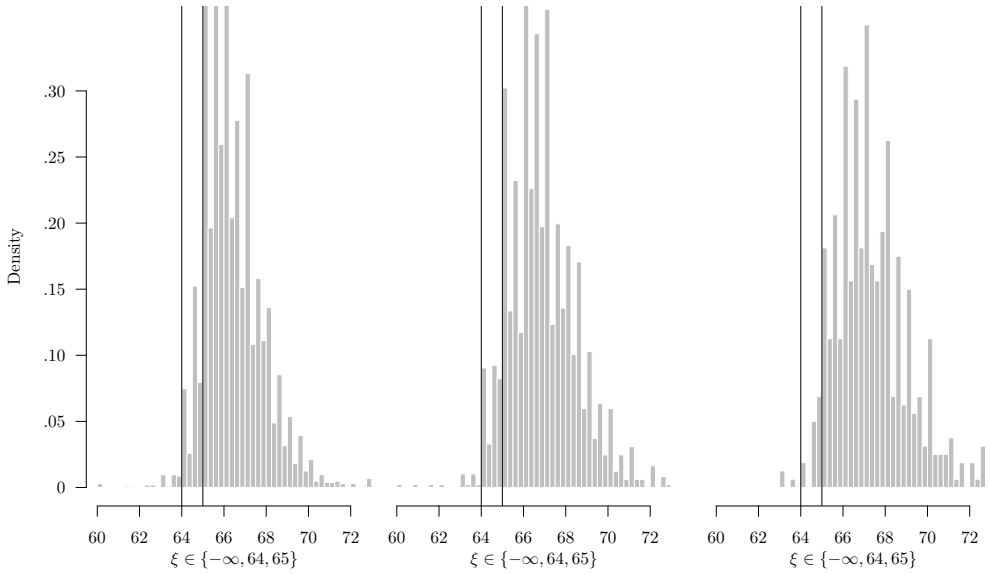

Figure 27: Birth cohort 1850-54.

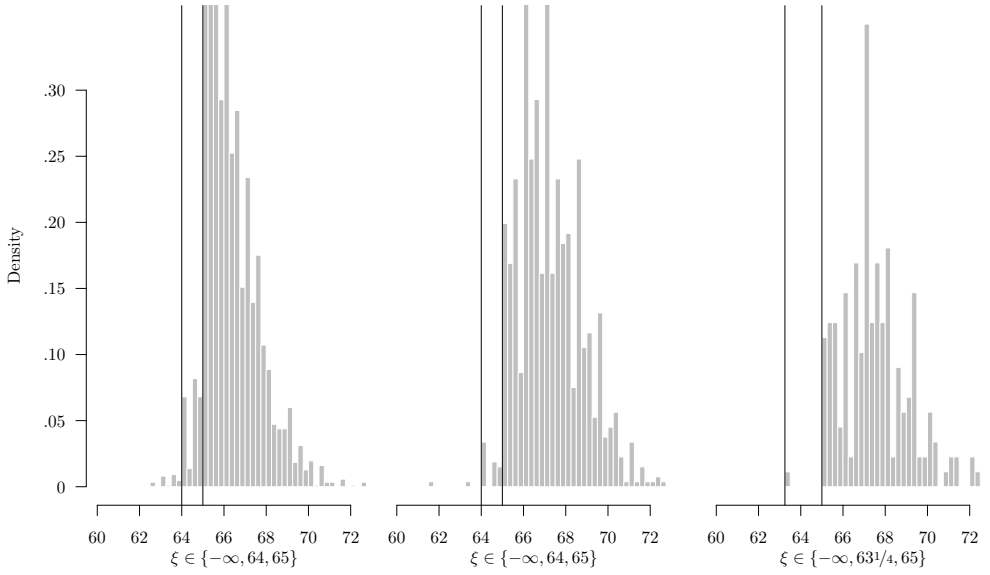

Figure 28: Birth cohort 1855-59.

European Historical Economics Society

EHES Working Paper Series

Recent EHES Working Papers

2018

EHES 141 Financial intermediation cost, rents, and productivity: An international comparison
Guillaume Bazot

EHES 140 The introduction of serfdom and labour markets
Peter Sandholt Jensen, Cristina Victoria Radu, Battista Severgnini and Paul Sharp

EHES 139 Two stories, one fate: Age-heaping and literacy in Spain, 1877-1930
Alfonso Díez-Minguela, Julio Martinez-Galarraga and Daniel A. Tirado-Fabregat

EHES 138 Two Worlds of Female Labour: Gender Wage Inequality in Western Europe, 1300-1800
Alexandra M. de Pleijt and Jan Luiten van Zanden

EHES 137 From Convergence to Divergence: Portuguese Economic Growth
Nuno Palma and Jaime Reis

EHES 136 The Big Bang: Stock Market Capitalization in the Long Run
Dmitry Kuvshinov and Kaspar Zimmermann

EHES 135 The Great Moderation of Grain Price Volatility: Market Integration vs. Climate Change, Germany, 1650–1790
Hakon Albers, Ulrich Pfister and Martin Uebele

EHES 134 The age of mass migration in Latin America
Blanca Sánchez-Alonso

EHES 133 Gravity and Migration before Railways: Evidence from Parisian Prostitutes and Revolutionaries
Morgan Kelly and Cormac Ó Gráda

All papers may be downloaded free of charge from: www.ehes.org

The European Historical Economics Society is concerned with advancing education in European economic history through study of European economies and economic history. The society is registered with the Charity Commissioners of England and Wales number: 1052680