THE AUSTRALIAN NATIONAL UNIVERSITY

WORKING PAPERS IN ECONOMICS AND ECONOMETRICS

Nonparametric Density Estimation for Stratified Samples

Robert Breunig
The Australian National University[*]

Working Paper No. 459

First version: February, 2001
Current version: November, 2005

ISBN: 086831 459 5

[*] Center for Economic Policy Research, Economics Program, Research School of Social Sciences, The Australian National University, Canberra ACT 0200, AUSTRALIA
phone: +61 (0)2 6125-2148; fax: +61 (0)2 6125-0087, e-mail: Robert.Breunig@anu.edu.au

Nonparametric Density Estimation for Stratified Samples

Robert Breunig
The Australian National University[*]

First version: February, 2001
Current version: October 17, 2006

Abstract

In this paper, we consider the non-parametric, kernel estimate of the density, $f(x)$, for data drawn from stratified samples. Much of the data used by social scientists is gathered in some type of complex survey violating the usual assumptions of independently and identically distributed data. Such effects induced by the survey structure are rarely considered in the literature on non-parametric density estimation, yet they may have serious consequences for our analysis, as shown in this paper.

A weighted estimator is developed which provides asymptotically unbiased density estimation for stratified samples. A data-based method for choosing the optimal bandwidth is suggested, using information on within-stratum variances and means. The weighted estimator and proposed bandwidth are shown to give smaller mean squared error for stratified samples than an un-weighted estimator and a commonly used method of choosing the bandwidth. Surprisingly, the single bandwidth outperforms optimally choosing stratum-specific bandwidths in some cases. Several illustrations from simulation are provided. We also show that the optimal sampling scheme in this case is always stratified sampling proportional to size, irrespective of the stratum-specific densities.

Keywords: *nonparametric density estimation, bandwidth selection, stratified sampling, optimal sampling*

JEL Classification: C14, C42

[*] School of Economics, Canberra ACT 0200, AUSTRALIA
phone: +61 (02) 6125-2148; fax: +61 (02) 6125-0087, e-mail: Robert.Breunig@anu.edu.au
An anonymous referee has pointed out several errors in a previous version of this working paper and I am very appreciative of those comments which allowed me to correct the current version of the paper. I am grateful to Aman Ullah for his suggestions and comments on this paper. I have also benefitted from conversations with Chris Skeels and Nilanjana Roy and the comments of participants in seminars at Georgia State University, the U.S. Bureau of Labor Statistics, University of California, Riverside, the Econometric Society Australasian Meetings 1998, University of Sydney and the University of New South Wales.

1 Introduction

The properties of kernel-based non-parametric density estimation are well-known for independent and identically distributed (i.i.d.) data[1]. Their use has become quite common in the social sciences for both descriptive and analytic purposes. In economics, density estimation has become a common part of the tools that analysts use to examine distributions of income, education, wages, consumption, receipt of government benefits, and many other variables. The advantages of using smooth non-parametric techniques over histrograms or parametric techniques to describe such distributions have become widely accepted.

The standard approach to density estimation in applied work in economics uses the i.i.d. assumption despite using survey data which violate that assumption. The data sets which economists and other social scientists use are typically generated using some type of complex sampling design.

Stratified sampling is probably the most commonly encountered sampling design in data used by applied social scientists. (Details of stratified sampling and more complex sampling schemes may be found in the econometric literature in Pudney (1989), Deaton (1997), Breunig and Ullah (1998) or in one of the traditional statistics texts such as Kish (1965) or Thompson (1992).) Though stratified sampling may be quite complicated in application, the primary effect of such sampling is that the population elements enter the sample with unequal probabilities. Therefore, in order to estimate model parameters we need to account for these unequal sampling probabilities.

The purpose of this paper is to begin the task of developing non-parametric density estimation for stratified survey data. Breunig (2001) extends the non-parametric kernel estimator to clustered data and demonstrates the large pointwise bias which results from ignoring the clustering in the data. This paper complements that work by considering stratified sampling.

In the case where large samples are available from each stratum, the natural approach would be to estimate stratum-specific densities and sum those using population proportions. We assume that it is not possible to do this in what follows. (Although in the numerical illustrations below, we provide comparisons

[1] Rosenblatt (1956) and Parzen (1962) are generally cited as the initiators of nonparametric density estimation. For summaries of the non-parametric literature and subsequent developments, see Silverman (1986), Härdle (1990), and Pagan and Ullah (1999).

to this 'optimal' approach.) Our approach is motivated by the fact that it is often the case that some strata are represented by few observations and it is not possible to efficiently estimate a separate density for those strata. In such cases, the distribution in the separate strata are not of independent interest. What is of interest is a population estimate of the density that uses the sampling information. Another possibility, and one that economists frequently face, is that of having some knowledge of the different strata from which the sample was drawn but no knowledge of which observation comes from which strata. Due to data confidentiality rules, analysts are not given such information, but are given weights which, at least in part, arise from the stratification in sampling. The approach suggested below allows for use of such weights even in the absence of knowledge about which observations belong to which strata.

After developing the simple model, we proceed to the main results: development of a weighted non-parametric density estimator which is asymptotically unbiased for stratified samples. Incorporating the sampling information into the choice of bandwidth selection, we provide the optimal bandwidth for the case where the data in each stratum are normally distributed. We examine the properties of the proposed bandwidth numerically and through simulation. We then derive the optimal sampling allocation for the stratified density estimator. Surprisingly, it differs from the optimal allocation for estimation of the mean.

2 Nonparametric Density Estimation: Stratified Sampling

Let us consider density estimation for data chosen under stratified sampling. Consider the following population model,

$$Y_{ij}\,, \qquad i = 1, \ldots, M \qquad j = 1, \ldots, N_i\,.$$

The total number of elements in the population is $\sum N_i = N$ and the proportion of elements in each stratum, i, is $\theta_i = \frac{N_i}{N}$. We treat the finite population within each stratum as large enough to be well approximated by a continuous distribution g_i, with mean μ_i and variance σ_i^2. We will only restrict these densities by the requirement that the first two moments exist and are finite for each stratum.

The distribution of interest is that of the finite population given by

$$f(Y) = \sum_{i=1}^{M} \theta_i g_i .^2 \tag{1}$$

Now consider a sample, where n_i elements (labelled y_{ij}, $j = 1, \ldots, n_i$) are drawn by simple random sampling with replacement independently from each stratum (i.e. a stratified sample). The total sample size is $\sum n_i = n$. The n_i may or may not be equal. Since both the n_i and the θ_i may vary, the sample inclusion probabilities are not equal for all elements in the sample. They will however, be equal for all elements in the same stratum. The probability that the j-th element in the i-th stratum is included in the sample is $\pi_{ij} = \pi_i = \frac{n_i}{N_i}$.

Rosenblatt's (1956) kernel estimator for the density in the ith stratum, based on the sample of size n_i may be written as

$$\widehat{g}_i(y) = \frac{1}{hn_i} \sum_{j=1}^{n_i} K_j \tag{2}$$

where $K_j = K(\frac{y_{ij}-y}{h})$ is a kernel function. In what follows, we assume that it is not possible or practical to generate a separate estimate of the distribution in each stratum. This may be because some strata have very small sample sizes or it may be that while sampling weights are available to the practitioner, information about which observation belongs to which strata is not.

Using the sample data from all strata the usual estimator for the density at a point y is

$$\hat{f}(y) = \frac{1}{nh} \sum_{i=1}^{M} \sum_{j=1}^{n_i} K\left(\frac{y_{ij}-y}{h}\right) = \sum_{i=1}^{M} \frac{n_i}{n} \widehat{g}_i(y) \tag{3}$$

where h is the window width, and the kernel $K(\cdot)$ is a symmetric function which satisfies:

(A1) (i) $\int K(\psi) d\psi = 1$

(ii) $\int \psi K(\psi) d\psi = 0$

(iii) $\int \psi^2 K(\psi) d\psi = \gamma_2 < \infty$

[2] Another way of thinking of this is that each stratum is an i.i.d. draw from a different superpopulation and that characteristics of the finite population are of interest for analysis. In this case, combining the super-population parameters using the stratum population proportions produces an overall density estimate of interest.

The estimator (3) of the population density, $f(y)$, is thus a sample-weighted average of the density estimates for each stratum. This estimator will not be unbiased for the parameter of interest. To see this, we write

$$E\hat{f}(y) = \sum_{i=1}^{M} \frac{n_i}{n} E\widehat{g}_i(y) \tag{4}$$

and by Taylor's series expansion

$$E\widehat{g}_i(y) = g_i(y) + bias_i(h) \tag{5}$$

where $bias_i(h)$ represents bias terms which will depend upon h. This provides

$$E\hat{f}(y) = \sum_{i=1}^{M} \frac{n_i}{n} g_i(y) + \sum_{i=1}^{M} \frac{n_i}{n} bias_i(h). \tag{6}$$

Usually we choose h such that $h \to 0$ as $n_i \to \infty$, therefore the $bias_i(h)$ terms will become small as the n_i become large. Even then, however, we still have bias arising from the fact that we are implicitly weighting the stratum-specific densities by the sample proportions.

It is thus clear that the density estimate, $\hat{f}(y)$, will only be asymptotically unbiased for (1) when

$$(i) \qquad \frac{n_i}{N_i} = \frac{n}{N}$$

or

$$(ii) \qquad g_i = g \qquad \forall i. \tag{7}$$

These conditions are unlikely to be met in most surveys. It is a common feature of surveys that sampling is disproportionate, violating condition (i). Even when the original survey design is such that the sample inclusion probabilities are equal in all strata, varying rates of non-response and other factors usually make the sampling disproportionate. This is often a desired trait when particular populations of interest are sampled more heavily relative to the rest of the population (Survey of Income and Program Participation (1991) for example) or when cost restricts sampling. (i.e. the case of Living Standards Measurement Study data from the World Bank where lower cost of sampling in urban areas leads to higher sampling proportions in these areas.) Though we are interested in an overall estimate of the density, it is problematic to assume that variables of interest will be identically distributed in different strata. Ignoring either this dis-proportionality in the survey design or the differences between strata will lead to biased estimation, even in the simple case of non-parametric kernel density estimation.

The solution is a weighted estimator

$$\hat{f}_w(y) = \frac{1}{h\sum w_i} \sum_{i=1}^{M} \sum_{j=1}^{n_i} w_i K\left(\frac{y_{ij}-y}{h}\right) \tag{8}$$

where $w_i \alpha \frac{N_i}{n_i}$. We set the weights proportional to the inverse of the selection probabilities. If we further require that $\sum w_i = 1$, then $w_i = \frac{N_i}{Nn_i}$. Then

$$\hat{f}_w(y) = \sum_{i=1}^{M} \frac{N_i}{N} \widehat{g}_i(y) = \sum_{i=1}^{M} \theta_i \widehat{g}_i(y) \tag{9}$$

As noted above, however, this is not unbiased for (1) since $\widehat{g}_i(y)$ is not unbiased for g_i. This bias will depend upon the choice of window width, h. Writing $bias_i$ for $bias_i(h)$, we can write

$$\hat{f}_w(y) = \sum_{i=1}^{M} \theta_i (g_i + bias_i) \tag{10}$$

and

$$bias\left(\hat{f}_w(y)\right) = \sum_{i=1}^{M} \theta_i bias_i \tag{11}$$

where the typical bias term upto $O(h^2)$ will depend on the second derivative of the true underlying density

$$bias_i = \frac{h^2}{2} g_i'' \gamma_2. \tag{12}$$

Assuming that the sampling is independent between strata (which is usually the case), we can also write

$$Var\left(\hat{f}_w(y)\right) = \theta_1^2 var\left(\widehat{g}_1\right) + \theta_2^2 var\left(\widehat{g}_2\right) + ... + \theta_M^2 var\left(\widehat{g}_M\right) \tag{13}$$

and upto $O(\frac{1}{nh})$

$$\int Var\left(\hat{f}_w(y)\right) dy = \frac{1}{h}\left[\int (K(\psi))^2 \, d\psi\right] \sum_{i=1}^{M} \frac{\theta_i^2}{n_i}. \tag{14}$$

Silverman (1986) provides details of the non-stratified case for sampling with replacement. If we consider each stratum as such a sample, it is then straightforward to work out (10) through (14).

Proposition 1: If the densities of strata 1 through M are given as g_1 through g_M, the population density $f(y)$ is estimated using a kernel density satisfying (A1), and a stratified sample of data is drawn independently in each stratum, then the window width which minimizes the mean-squared error of $\widehat{f}_w(y)$ will be

$$h_{st} = \left(\gamma_2^2\right)^{-\frac{1}{5}} \left(\left[\int\limits_{\psi} \left(K(\psi)\right)^2 d\psi \right] \right)^{\frac{1}{5}} \left(\sum_{i=1}^{M} \frac{\theta_i^2}{n_i} \right)^{\frac{1}{5}} \left(\int\limits_{y} \left[\sum_{i=1}^{M} \theta_i \left(g_i''\right) \right]^2 dy \right)^{-\frac{1}{5}}. \tag{15}$$

Proof: using (10) through (14) we write the mean squared error of $\hat{f}_w(y)$ as $Var\left(\hat{f}_w(y)\right) + \left(bias\left(\hat{f}_w(y)\right)\right)^2$. The integrated mean squared error is then

$$\int\limits_{y} \left\{ Var\left(\hat{f}_w(y)\right) + \left(bias\left(\hat{f}_w(y)\right)\right)^2 \right\} dy.$$

We minimize this expression with respect to h to get the result in Proposition 1.$\diamondsuit$

In order to implement this result, we need to know the second derivative of the true underlying density. Of course, this will normally not be available. One solution to this problem is to specify a family of distributions which will allow a value to be assigned to the term $\int\limits_{y} \left[\sum_{i=1}^{M} \theta_i \left(g_i''\right)\right]^2 dy$ in (15). For the i.i.d. case, it has been shown that when $f(y)$ is normally distributed that the optimal window width will be h*=$1.06\sigma n^{-\frac{1}{5}}$ where σ is the standard deviation of y. This choice of window width is commonly employed in econometrics software packages (Stata Corporation (2005), for example) and is a frequently used starting point for other bandwidth selection techniques such as cross-validation.

Here it is natural to ask whether a similar reference window width can be derived based upon underlying normal distributions in all of the strata. Corollary 1 gives the value of that reference window width.

Corollary 1: If g_1 through g_M are normally distributed with mean μ_i and variance σ_i^2 and the density is estimated using a standard normal kernel, then the optimal window width (in the mean squared error sense) will be

$$h_{st} = 0.87 \left(\sum_{i=1}^{M} \frac{\theta_i^2}{n_i} \right)^{\frac{1}{5}} \left(\lambda_1 + \lambda_2\right)^{-\frac{1}{5}} \tag{16}$$

where λ_1 is a weighted sum of stratum-specific standard deviations

$$\lambda_1 = \frac{3}{8}\sum_{i=1}^{M} \theta_i^2 \sigma_i^{-5}$$

and λ_2 is a weighted sum of a function of the distance between stratum means

$$\lambda_2 = \sum_{i=1}^{M}\sum_{l \neq i}^{M} \theta_i \theta_l \frac{\left(\sigma_i^2+\sigma_l^2\right)^{-\frac{5}{2}}}{\sqrt{2}} \left\{3 - 6\frac{(\mu_i-\mu_l)^2}{(\sigma_i^2+\sigma_l^2)} + \frac{(\mu_i-\mu_l)^4}{(\sigma_i^2+\sigma_l^2)^2}\right\} e^{-\frac{1}{2}\frac{(\mu_i-\mu_l)^2}{\left(\sigma_i^2+\sigma_l^2\right)}}$$

Proof: For the case of a standard normal kernel $\gamma_2^2 = 1$ and $\int\limits_{\psi} (K(\psi))^2 \, d\psi = \frac{1}{2\sqrt{\pi}}$. We can write

$$\int\limits_{y} \left[\sum_{i=1}^{M} \theta_i \left(g_i''\right)\right]^2 dy = \int\limits_{y} \sum_{i=1}^{M} \theta_i^2 \left(g_i''\right)^2 dy \; + \int\limits_{y} \sum_{i=1}^{M}\sum_{l \neq i}^{M} \theta_l \theta_i \left(g_i''\right)\left(g_l''\right) dy$$

and for normal densities replace g_i'' with $\frac{1}{\sigma_i^3\sqrt{2\pi}}\left[1 - \left(\frac{y-\mu_i}{\sigma_i}\right)^2\right] e^{-\frac{1}{2}\left(\frac{y-\mu_i}{\sigma_i}\right)^2}$. Then the first term, $\int\limits_{y} \sum_{i=1}^{M} \theta_i^2 \left(g_i''\right)^2$, becomes $\frac{3}{8\sqrt{\pi}} \sum_{i=1}^{M} \sigma_i^{-5}\theta_i^2$. The second term can be calculated by integrating the product of g_i'' and g_j''. Using these results, calculate λ_1 and λ_2 and replace in the formula for h_{st}.$\diamondsuit$

We note that the optimal window width is inversely proportional to a weighted sum of the strata sample sizes, n_i. In the case where $n_i = \frac{n}{M}$ and $\theta_i = \frac{1}{M}$, then $\sum_{i=1}^{M} \frac{\theta_i^2}{n_i} = n$ and the window width will be proportional to $n^{-\frac{1}{5}}$ as in the non-stratified case, but the proportionality constant will differ from the usual 1.06σ. When strata share common means and variances, and the population and sample proportions are equal in all strata, this result collapses to the usual optimal window width for normal density: $h^* = 1.06\sigma n^{-\frac{1}{5}}$. Note that however if $\sigma_i = \sigma$ for all strata, that the population standard deviation may still differ from σ and $h_{st} \neq 1.06\sigma n^{-\frac{1}{5}}$.

When strata share common means and variances, $h_{st} = 1.06\sigma \left(\sum_{i=1}^{M} \frac{\theta_i^2}{n_i}\right)^{\frac{1}{5}}$. Thus even in the case of homogeneous populations in all strata, the optimal window width is different than the usual h^* unless $\theta_i = \frac{n_i}{n}$. This is analagous to the case of estimation of the mean, where even when all strata have identical means, the variance of the estimator $\overline{y}$ is different for a stratified sample than for a simple random sample.

In practice we can replace σ_i with some consistent estimator like $\sqrt{s_i^2}$ and μ_i with its estimate, $\overline{y}$.

2.0.1 Numerical Properties

In this section, we compare the integrated mean squared error under different choices for the window width. For the case where we use only one window width and estimate the density using the entire sample of data, we compare the optimally chosen window width, h_{st}, derived above and h^*, the standard reference window width for the i.i.d. case. We consider the simplest case of two strata, both of which are normally distributed. We also consider separate estimation of the two strata using an optimally chosen window width for the $j-th$ strata ($h_j^* = 1.06 * \sigma_j n_j^{-1/5}$) and combining the stratum-specific density estimates using (9). This exercise is meant to be an illustration of the trade-offs involved in these options rather than a compelling practical example. In practice, the more interesting case is that of many strata with very small sample sizes in each stratum.

Let the two strata populations be normally distributed with means μ_1 and μ_2 and standard deviations σ_1 and σ_2. In Figure 1a, we see the effects on the window width as we vary the means of the two strata, holding the standard deviations constant. Figure1b shows the effect on the various window width choices when we hold the stratum means constant and allow the standard deviations of the two strata to become increasingly different.

For the case of two strata, we can write the window width in (16) as a function of the difference in means, $\phi = \mu_2 - \mu_1$ and the difference in standard deviations, $\gamma = \frac{\sigma_2}{\sigma_1}$. In that case, λ_1 and λ_2 become

$$\lambda_1 = \frac{3}{8\sigma_1^5}\left[\theta_1^2 + \gamma^{-5}\theta_2^2\right]$$

and

$$\lambda_2 = \sqrt{2}\theta_1\theta_2\left(\sigma_1^2\left(1+\gamma^2\right)\right)^{-\frac{5}{2}}\left\{3 - 6\frac{\phi^2}{\sigma_1^2\left(1+\gamma^2\right)} + \frac{\phi^4}{\left(\sigma_1^2\left(1+\gamma^2\right)\right)^2}\right\}e^{-\frac{1}{2}\frac{\phi^2}{\sigma_1^2(1+\gamma^2)}}$$

For the purpose of the illustration, we fix $\theta_1 = \theta_2 = \frac{1}{2}$, $n_1 = n_2 = 50$, and $\sigma_1^2 = 1$. In Figure 1a, we can see that as the difference between strata means increases, h^* grows without bound. h_{st} on the other hand, increases for a period, but will asymptotically approach .4848. (The value of h_1^* and h_2^*, $1.06n^{-1/5}$, is $\lim_{\phi\to\infty} h_{st}$ for this set of parameter values.) Intuitively, the optimal estimator, h_{st}, increases when the combined strata are unimodal, but once the means are far enough apart for the density to exhibit bi-modality, h_{st} begins to decrease. The effect of this is that the density estimation is essentially being conducted separately on each stratum, the small window width giving near zero weight to comparisons between elements in different strata. The stratum-specific window widths are equal since they only depend upon the (identical) stratum-specific variances. Figure 1b provides the same illustration for strata with identical means, but increasingly different standard deviations. Again, h_{st}, is not going to grow without bound because it takes into account the fact that the increasing sample variation is the result of two strata with two different underlying distributions. (It is possible to show that $\lim_{\gamma \to \infty} h_{st} \approx .556$.)

Table 1 presents the values of h^* and h_{st} at various points from Figure 1. The second last column of Table 1 gives the ratio of the approximate IMSE (upto $O(\frac{1}{nh})$) of $\widehat{f_w}(y)$ using a standard normal kernel and employing both h^* and h_{st}. We compare their ratio as a measure of the efficiency loss of using h^*. The last column compares the 'ideal' approach of using stratum-specific window widths and a weighted combination of stratum-specific densities as in (9). Here the integrated mean squared error is numerically calculated using

$$\begin{aligned} IMSE\left[\widehat{f_w}(y)\right] &= \int\limits_x \left\{ Var\left(\hat{f}_w(y)\right) + \left(bias\left(\hat{f}_w(y)\right)\right)^2 \right\} dx \\ &= \int\limits_x \left\{ \frac{1}{h}\left[\int (K(\psi))^2 \, d\psi\right] \sum_{i=1}^{M} \frac{\theta_i^2}{n_i} + \frac{h^4}{4}\gamma_2^2 \left[\sum_{i=1}^{M} \theta_i g_i''\right]^2 \right\} dx \end{aligned}$$

and the appropriate values for h and the other variables based upon a standard normal kernel, and two normal densities with $\mu_1 = 0$, $\sigma_1 = 1$, and mean and standard deviation of the second stratum as specified. We calculate the variance and mean of the mixture of normals for h^* using standard formulas such as found in Behboodian (1970). Table 1a compares the loss of efficiency for two strata with equal standard deviations as the difference between strata means increases. In Table 1b, the means are held constant while strata standard deviations vary.

Given the use of the weighted estimator for the density, using the proper window width gives large improvements in mean squared error over the standard reference window width. This is true even when the sampling is proportional (to stratum size). As the two stratums become increasingly different (either in mean or in standard deviation) the gains in integrated mean squared error become quite large. When we compare the optimal (under the constraint of using only one parameter) window width h_{st} with the 'ideal' approach of separate strata-specific density estimation, we see that there is little loss in mean squared error from using one window width in the case where the different strata have identical standard deviations. As the distance between the strata increases (where $\mu_2 - \mu_1$ is between 1 and 3.5), using h_{st} actually provides superior integrated mean squared error, but the gains are small. The improvement in variance from a larger window width under h_{st} must dominate any bias penalty in these cases. Examining the point-wise data, the larger window width of h_{st} does a better job of estimating the density between the two modes where there is mixing of the two distributions. In the case where stratum-specific standard deviations differ, however, the constraint of using only one window width, even when chosen optimally, comes at fairly high penalty in terms of integrated mean squared error.

When the difference between means is greater than 2, using h^* results in very large efficiency losses compared to using h_{st}. This corresponends to the results presented in the simulation below. For the case of dis-proportionate sampling, the relative loss of IMSE is not much different than in the case of proportional sampling. As we will see from the simulation, however, the bias is much greater using h^*. Since the efficiency measure considered here includes integrated bias, it is perhaps not a good measure of the pointwise bias from using h^*. However, it is quite clear from the figures presented in the simulation exercise below that this pointwise bias will be unacceptably large.

Figures 2 and 3 provide a graphical depiction of the ratio of integrated mean squared errors from the last two columns of Table 1a and Table 1b.

2.0.2 Simulation study

In a simulation study using a similar simple set-up, we consider the proposed optimal window width, h_{st} versus $h^* = 1.06\sigma n^{-\frac{1}{5}}$ and $h_a = .9\text{Min}(\sigma, \frac{inter-quartile\ range}{1.34})n^{-1/5}$. We also include comparison with estimation of stratum-specific densities which are then combined using (9), as described above. h_a has been shown to be superior to h^* for mixtures of normals and bimodal densities, see Silverman (1986). For clarity, we consider sampling from two strata, where the population in each strata is equal and the underlying densities are normal with mean μ_i and standard deviation σ_i. For the simulation, we fix $n_1 = 50$, $\mu_1 = 0$ and $\sigma_1 = 1$ while varying the sample size, the mean, and the standard deviation of stratum 2 only. Proportional sampling thus implies $\frac{n_2}{n_1} = 1$, otherwise the sampling is disproportionate.

For proportional sampling, we consider the benchmark case when there is no difference in mean or standard deviation between the two strata. We then consider how estimation changes using the proposed h_{st} as the difference between the two strata means increases, as the difference between the standard deviations increases, and as both change. We then consider the same cases for disproportionate sampling.

We conduct 1000 repetitions for each case. Each repetition involves drawing a sample from the two strata, estimating the five candidate window widths (h^*, h_a, h_1^*, h_2^* and h_{st}) based upon the formulas above (with σ replaced by s, and μ by $\overline{y}$), and estimating the non-parametric density at 800 points. The figures give the average estimate of the density over the 1000 repetitions. Tables 2 presents the average calculated window widths and the various combinations which have been considered in the simulation exercise. Table 3 presents our numerically estimated value of the integrated mean squared error. These differ from the theoretical ones above because of the order of approximation used in the theoretical calculations.

We first consider the case of proportional sampling ($n_1 = n_2 = 50$) when both strata have mean zero and variance one. The average h_{st} is the same as the average h^* and the two average density estimates are identical. This is as expected given the discussion above. In this case, the stratification is spurious since the two strata are exactly identical. Note that in practice, however, the density estimate using h^* will be superior to that using h_{st} since calculation of h_{st} involves computing two strata means and two strata standard deviations instead of one total sample standard deviation. The estimation of four quantities instead of one introduces more variability into the estimate of h_{st} than h^*.

The improvement provided by using h_{st} as opposed to h^* or h_a is dramatic when the difference between the strata means grows and the overall density becomes bi-modal. As can be seen in Figure 3, h^* tends to oversmooth the peaks. h_a gives improved performance and reduces this over-smoothing, but h_{st} can be seen to match the peaks even better than either h^* or h_a. When means between strata are equal, but variances differ, the same results hold: h_a improves performance over h^*, but h_{st} matches the density better than either.

As noted above, proportional sampling will tend to be the exception in most cross-sectional data sets used by economists. The proposed optimal window width, h_{st} combined with the weighted density estimator of (8), proves to be a very powerful tool for non-proportional sampling. This is examined in the remaining figures.

When the sampling is not proportionate and the strata differ in either means or variances, the unweighted estimator will be biased as discussed above. This is clear from Figures 9 and 10 where we compare density estimation for two strata with equal standard deviations but different means. In both cases, the weighted estimator using h_{st} clearly outperforms unweighted estimation with any (unique) window width. (We present, therefore, only the comparison between weighted estimation using h_{st} and unweighted density estimation using h^*.) Here stratum 2 is sampled twice as intensively as stratum 1, thus the elements from stratum 2 receive a weight that is half that of elements in stratum 1. We would point out that this is not a particularly large difference in weights. In many of the data sets used in applied economic work, the sampling disproportion is greater than 10 between certain strata, so the results from ignoring the weighting in this case will be even more dramatic with even larger resulting bias.

Figures 11 and 12 illustrate the case of equal strata means and different variances and the case of variation between strata of both means and standard deviations. Again, the same results hold. Large bias is incurred by ignoring the structure of the sampling.

2.0.3 Optimal allocation

If we have some information about the means and standard deviations in the various strata (perhaps from a previous survey), can we use that information to construct an optimal sampling allocation to minimize the integrated mean squared error of the estimator of $f(y)$? We know that in the case of stratified sampling for mean estimation that oversampling (relative to population proportions) strata with higher variance can give a more precise estimate of the mean. Does a similar result hold here?

Curiously, it turns out that proportional sampling will be the optimal allocation in all cases, given that we are optimally choosing h_{st}.

Proposition 2: If the densities of strata 1 through M are given as g_1 through g_M, the population density $f(y)$ is estimated using a kernel density satisfying (A1), and the window width is chosen as (15), then the sampling allocation which minimizes the integrated mean squared error of $\hat{f}_w(y)$ is sampling proportional to stratum size,

$$n_i = n\theta_i.$$

Proof: The integrated mean squared error of $\hat{f}_w(y)$ is

$$\frac{1}{h}\left[\int (K(\psi))^2 \, d\psi\right] \sum_{i=1}^{M} \frac{\theta_i^2}{n_i} + \frac{h^4}{4}\gamma_2^2 \int_y \left[\sum_{i=1}^{M} \theta_i g_i''\right]^2 dy$$

Replacing h with h_{st} yields

$$\begin{aligned} IMSE(\hat{f}_w(y)) &= \frac{5}{4}\gamma_2^{\frac{2}{5}} \left[\int (K(\psi))^2 \, d\psi\right]^{\frac{4}{5}} \left(\int_y \left[\sum_{i=1}^{M} \theta_i g_i''\right]^2 dy\right)^{\frac{1}{5}} \left(\sum_{i=1}^{M} \frac{\theta_i^2}{n_i}\right)^{\frac{4}{5}} \\ &= k^*. \left(\sum_{i=1}^{M} \frac{\theta_i^2}{n_i}\right)^{\frac{4}{5}} \end{aligned}$$

If we minimize this quantity with respect to $n_1, ..., n_M$ constrained by $\sum n_i = n$, a typical equation will be

$$\frac{\partial IMSE}{\partial n_j} = -\frac{4}{5}k^* \left(\sum_{i=1}^{M} \frac{\theta_i^2}{n_i}\right)^{-\frac{1}{5}} \frac{\theta_j^2}{n_j^2} + \lambda = 0$$

where λ is the Lagrange multiplier. If we solve for λ, multiply both sides of the equation by n_j^2, take the square root of both sides of the equation, and solve for $\sqrt{\lambda}\sum n_j = \sqrt{\lambda}n = \left(\frac{4}{5}k^*\right)^{\frac{1}{2}} \left(\sum_{i=1}^{M} \frac{\theta_i^2}{n_i}\right)^{-\frac{1}{10}}$. Replacing λ in the above equation then provides $\frac{\theta_j^2}{n_j^2} = \frac{1}{n^2}$ and $n_j = n\theta_j$. $\diamondsuit$

This is a somewhat surprising result given the intuition from the mean estimation problem. However, in this case, we are not estimating any single point from each stratum, but instead the entire distribution. Even from a stratum whose distribution has a small variance we will need a sample size sufficiently large to estimate the contribution of that stratum to the overall population density.

3 Concluding Remarks

This paper is an attempt to begin unifying the literatures on survey design and nonparametric density estimation. As such we begin by analyzing the plug-in window width for normal data, the point of departure for most theoretical considerations of nonparametric density estimation as well as a useful bandwidth to generate a first guess at the distribution or for use as a starting point for other data-driven bandwidth selection techniques such as nearest neighbor and cross-validation. It is instructive to see how the standard results change when stratification is introduced.

The framework here is general and does not depend upon any minimal strata sample sizes. For the simple examples considered in the simulations, it may be that using a different bandwidth for each stratum, estimating individual stratum-specific densities and then combining them using (9) will provide an adequate alternative. However for cases where there are many strata, some with only a handful of elements, such a technique is not feasible. The technique presented in this paper, to choose one bandwidth for all the data which takes into account the strata differences, will work in this case. Of course, knowledge of stratum-specific means and variances (or access to reasonable estimates thereof) is necessary. This is a problem which is frequently faced by survey statisticians in designing an optimal allocation. Using pretests, previous survey samples, or simple aggregation rules to combine similar strata are all ways around this problem, though all are imperfect. The technique does not relieve the researcher of the need to make intelligent choices according to the particular application.

Many problems remain to be addressed. One problem for economists is the dearth of information on the survey design behind the data. Occasionally we know something about survey weights, rarely do we know which observations come from which particular strata or clusters. The technique presented here may be used in that case, provided some other source of information about stratum-specific means and variances is available. Lack of survey information remains an impediment to improving our analytical techniques and we need to make a more concerted effort to have such information included with data.

References

Behboodian, J. (1970). On a mixture of normal distributions. *Biometrika*, 57(1):215–217.

Breunig, R. (2001). Density estimation for clustered data. *Econometric Reviews*, 20(3):353–367.

Breunig, R. and Ullah, A. (1998). Econometric analysis in complex surveys. In Giles, D. and Ullah, A., editors, *Handbook of Applied Economic Statistics.* Klewer Academic Press.

Deaton, A. (1997). *The Analysis of Household Surveys: A Microeconometric Approach to Development Policy.* Johns Hopkins University Press, Baltimore, MD.

Härdle, W. (1990). *Applied Nonparametric Regression.* Cambridge University Press.

Kish, L. (1965). *Survey Sampling.* John Wiley & Sons, New York, NY.

Pagan, A. R. and Ullah, A. (1999). *Nonparametric Econometrics.* Cambridge University Press.

Parzen, E. (1962). On estimation of a probability density function and mode. *Annals of Mathematical Statistics*, 33:1965–1076.

Pudney, S. (1989). *Modelling Individual Choice: The Econometrics of Corners, Kinks, and Holes.* Basil Blackwell.

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

Silverman, B. (1986). *Density Estimation for Statistics and Data Analysis.* London: Chapman and Hall.

Stata Corporation (2005). *Stata User's Guide, Release 9.* College Station, TX: Stata Press.

Survey of Income and Program Participation (1991). *User's Guide.* Washington, DC.

Thompson, S. K. (1992). *Sampling.* John Wiley & Sons, New York, NY.

Table 1a: Comparison of window widths and integrated mean squared errors (IMSE) from weighted and unweighted estimation

Identical Standard Deviations

$\mu_2 - \mu_1$	h^*	h_{st}	$IMSE(h^*)$	$IMSE(h_{st})$	$\frac{IMSE(h_{st})}{IMSE(h^*)}$	$\frac{IMSE(h_{st})}{IMSE(h_1^*,h_2^*)}$
Proportional Sampling: n2=n1						
0.0	0.42199	0.42168	0.00836	0.00836	1.00000	1.14870
0.5	0.43498	0.43513	0.00810	0.00810	1.00000	1.11319
1.0	0.47180	0.47833	0.00737	0.00737	0.99963	1.01265
1.5	0.52749	0.55035	0.00643	0.00641	0.99656	0.88014
2.0	0.59679	0.58571	0.00602	0.00602	0.99928	0.82701
2.5	0.67552	0.53979	0.00738	0.00653	0.88512	0.89736
3.0	0.76076	0.49827	0.01140	0.00708	0.62081	0.97214
3.5	0.85056	0.47889	0.01797	0.00736	0.40974	1.01147
4.0	0.94361	0.47387	0.02639	0.00744	0.28199	1.02219
4.5	1.03904	0.47572	0.03645	0.00741	0.20335	1.01821
5.0	1.13625	0.47930	0.04896	0.00736	0.15028	1.01061
Non-proportional Sampling: $n2 = 2 * n1$						
0.0	0.38912	0.39811	0.00665	0.00664	0.99898	1.15927
0.5	0.40110	0.41080	0.00644	0.00644	0.99888	1.12344
1.0	0.43505	0.45159	0.00587	0.00586	0.99732	1.02197
1.5	0.48640	0.51958	0.00513	0.00509	0.99189	0.88824
2.0	0.55030	0.55296	0.00478	0.00478	0.99995	0.83462
2.5	0.62290	0.50961	0.00571	0.00519	0.90833	0.90561
3.0	0.70150	0.47041	0.00858	0.00562	0.65550	0.98108
3.5	0.78430	0.45212	0.01329	0.00585	0.44007	1.02078
4.0	0.87011	0.44738	0.01935	0.00591	0.30553	1.03160
4.5	0.95811	0.44912	0.02660	0.00589	0.22138	1.02758
5.0	1.04775	0.45250	0.03562	0.00584	0.16409	1.01991

$IMSE(h_1^*, h_2^*) = .00728$ for all rows in the top panel.

$IMSE(h_1^*, h_2^*) = .00573$ for all rows in the bottom panel.

Table 1b: Comparison of window widths and integrated mean squared errors (IMSE) from weighted and unweighted estimation

Identical Means

σ_2/σ_1	h^*	h_{st}	$IMSE(h^*)$	$IMSE(h_{st})$	$\frac{IMSE(h_{st})}{IMSE(h^*)}$	$\frac{IMSE(h_{st})}{IMSE(h_1^*,h_2^*)}$
Proportional Sampling: n2=n1						
1.0	0.42199	0.42168	0.00836	0.00836	1.00000	1.14870
1.5	0.53794	0.49888	0.00716	0.00707	0.98786	1.16513
2.0	0.66723	0.53353	0.00746	0.00661	0.88582	1.21051
2.5	0.80345	0.54689	0.00952	0.00645	0.67740	1.26530
3.0	0.94361	0.55208	0.01389	0.00639	0.45980	1.31608
3.5	1.08617	0.55426	0.02136	0.00636	0.29781	1.35946
4.0	1.23031	0.55526	0.03291	0.00635	0.19298	1.39578
4.5	1.37553	0.55575	0.04967	0.00634	0.12773	1.42623
5.0	1.52152	0.55602	0.07298	0.00634	0.08690	1.45196
5.5	1.66808	0.55616	0.10430	0.00634	0.06079	1.47391
6.0	1.81506	0.55625	0.14528	0.00634	0.04363	1.49281
Non-proportional Sampling: $n2 = 2 * n1$						
1.0	0.38912	0.39811	0.00665	0.00664	0.99898	1.15927
1.5	0.49604	0.47099	0.00565	0.00562	0.99437	1.11553
2.0	0.61526	0.50370	0.00578	0.00525	0.90896	1.12066
2.5	0.74087	0.51631	0.00720	0.00512	0.71153	1.14435
3.0	0.87011	0.52121	0.01031	0.00507	0.49199	1.17002
3.5	1.00157	0.52327	0.01568	0.00505	0.32233	1.19280
4.0	1.13448	0.52421	0.02400	0.00504	0.21022	1.21202
4.5	1.26839	0.52468	0.03610	0.00504	0.13963	1.22807
5.0	1.40301	0.52493	0.05293	0.00504	0.09519	1.24154
5.5	1.53815	0.52507	0.07556	0.00504	0.06666	1.25295
6.0	1.67368	0.52515	0.10518	0.00504	0.04788	1.26270

$IMSE(h_1^*, h_2^*) = .00728$ for first row of the top panel and decreases to .00425 in the last row of the top panel.

$IMSE(h_1^*, h_2^*) = .00573$ for first row of the bottom panel and decreases to .00399 in the last row of the bottom panel.

Table 2: Results of simulation exercise
Weighted Non-Parametric Density Estimation for Stratified Samples
Average window width values for simulations

μ_2	σ_2	n_2	h_{st}	h_a	h^*	h_1^*	h_2^*	Figure
0	1	50	0.41844	0.34562	0.42070	0.48247	0.48151	4
2	1	50	0.55608	0.50341	0.59581	0.48382	0.48444	5
3	1	50	0.48845	0.64586	0.76067	0.48140	0.48047	6
0	3	50	0.54803	0.57818	0.93639	0.48199	1.44148	7
3	3	50	0.55413	0.86537	1.13040	0.48174	1.44263	8
2	1	100	0.52969	0.45116	0.53613	0.48345	0.42115	9
3	1	100	0.46465	0.57394	0.67604	0.48102	0.42221	10
0	3	100	0.51793	0.65675	0.97623	0.48228	1.26164	11
3	3	100	0.52446	0.94106	1.12422	0.48304	1.26402	12

Table contains average results over 1000 simulations
Stratum 1 values are fixed at $n_1 = 50$, $\sigma_1 = 1$, and $\mu_1 = 0$.

Table 3: Results of simulation exercise
Weighted Non-Parametric Density Estimation for Stratified Samples
Average integrated mean squared error from the simulations

			Integrated Mean Squared Error			
μ_2	σ_2	n_2	h_{st}	h_a	h^*	h_1^* and h_2^*
0	1	50	0.00571	0.00648	0.00570	0.00575
2	1	50	0.00353	0.00374	0.00334	0.00419
3	1	50	0.00433	0.00419	0.00486	0.00453
0	3	50	0.00453	0.00460	0.00653	0.00328
3	3	50	0.00447	0.00562	0.00835	0.00315
2	1	100	0.00282	0.01085	0.00994	0.00329
3	1	100	0.00356	0.01430	0.01394	0.00366
0	3	100	0.00351	0.00911	0.01269	0.00302
3	3	100	0.00349	0.01401	0.01621	0.00280

Table contains average results over 1000 simulations
Stratum 1 values are fixed at $n_1 = 50$, $\sigma_1 = 1$, and $\mu_1 = 0$.

Figure1a

h^*, h_{st}, h_1^* and h_2^* for two strata with $\sigma_1=\sigma_2=1$

$n_1=n_2=50$

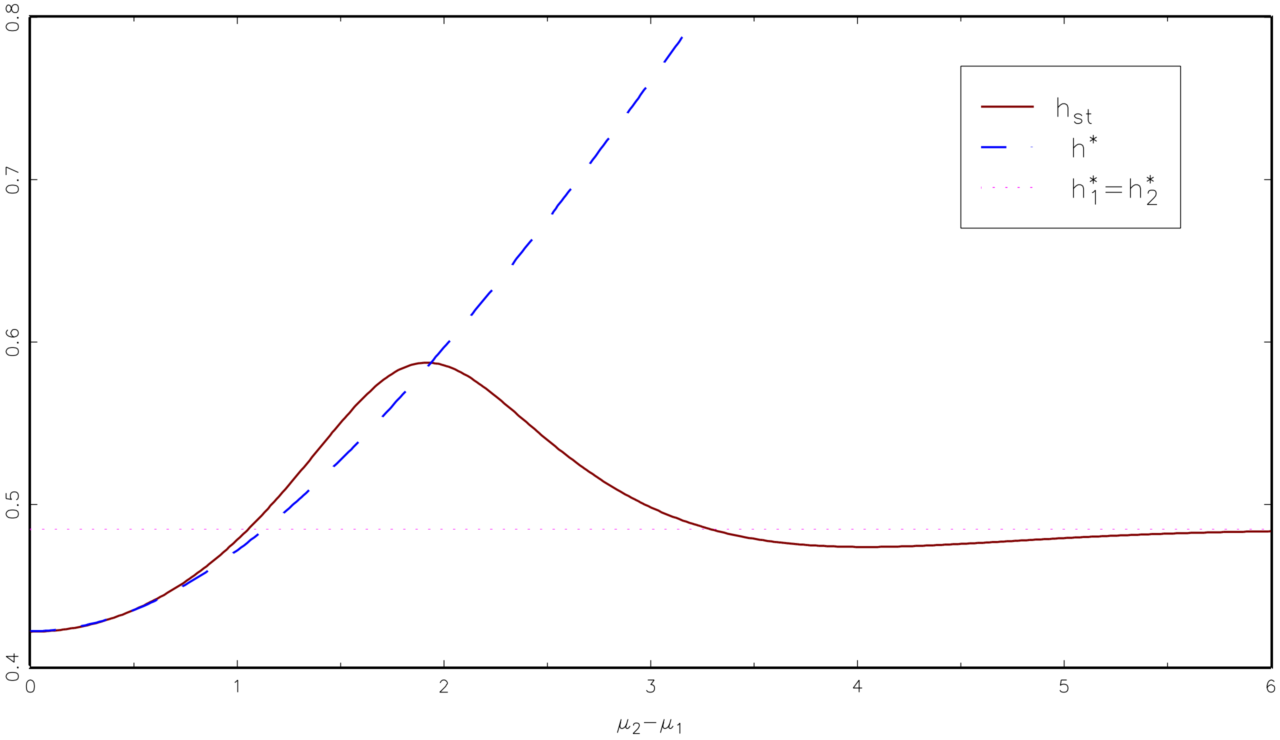


Figure1b

h^*, h_{st}, h_1^* and h_2^* for two strata with $\mu_1=\mu_2=0$

$n_1=n_2=50$

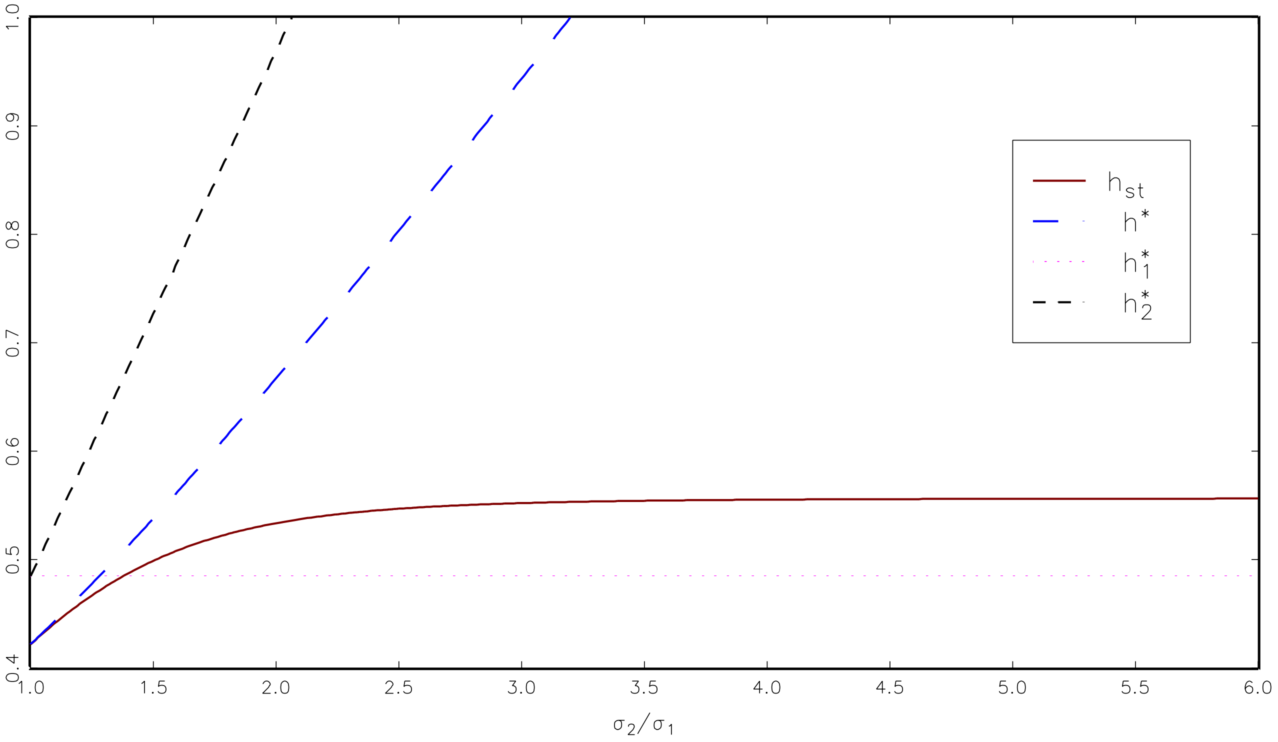


Figure2a

Ratio of Integrated Mean Squared Errors

$\sigma_1 = \sigma_2=1$ $n_1=n_2=50$

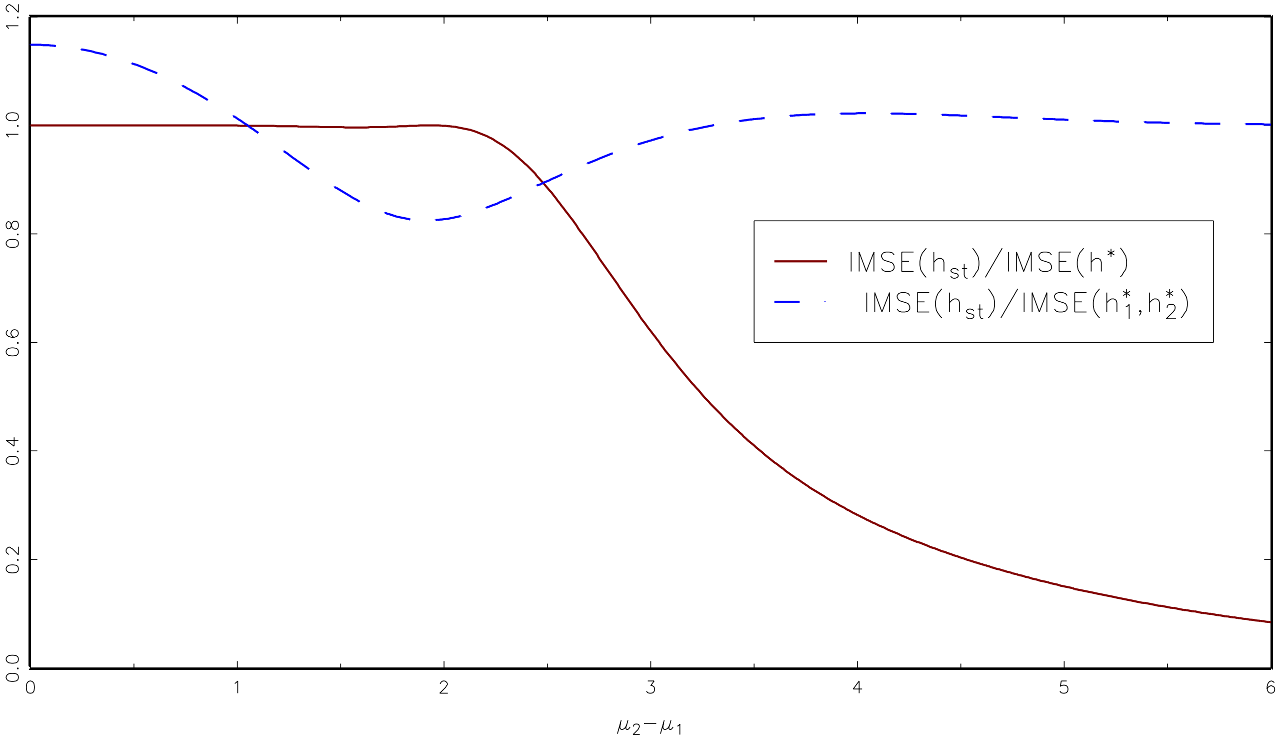


Figure2b

Ratio of Integrated Mean Squared Errors

$\mu_1=\mu_2=0$ $n_1=n_2=50$

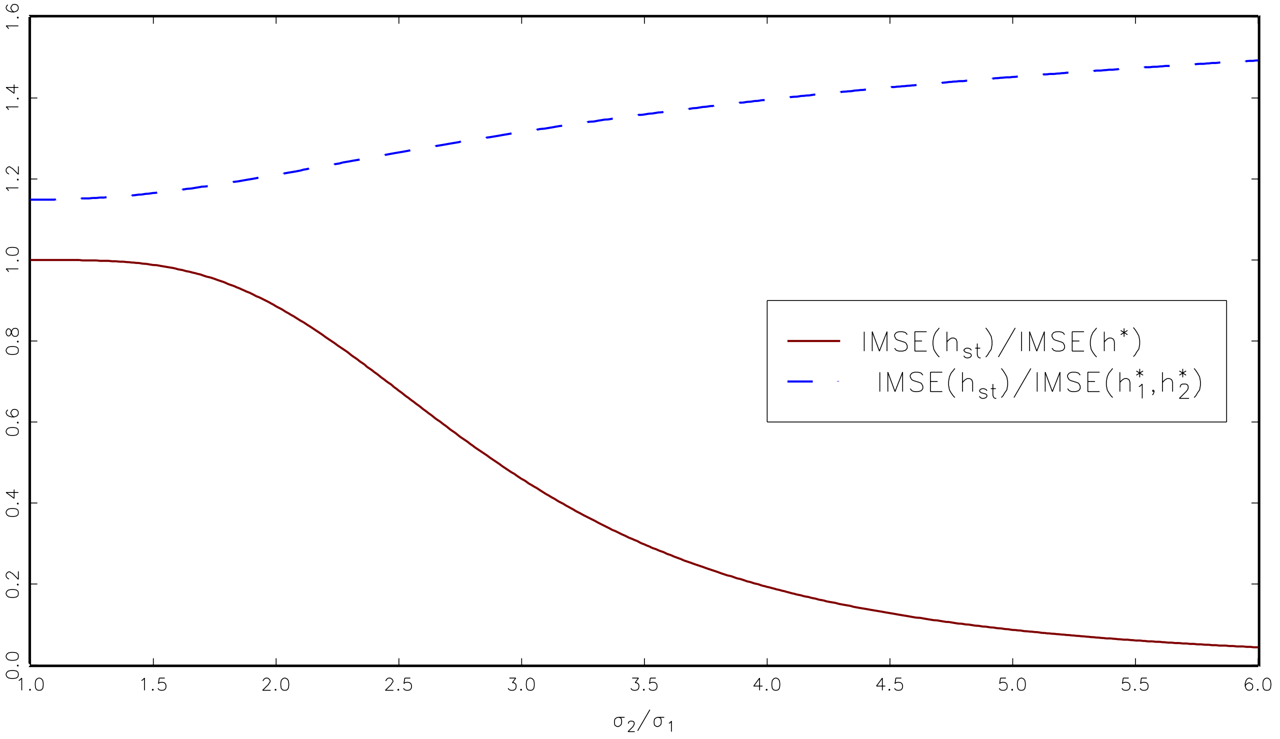


Figure3a

Ratio of Integrated Mean Squared Errors

$\sigma_1 = \sigma_2 = 1$ $n_1=50, n_2=100$

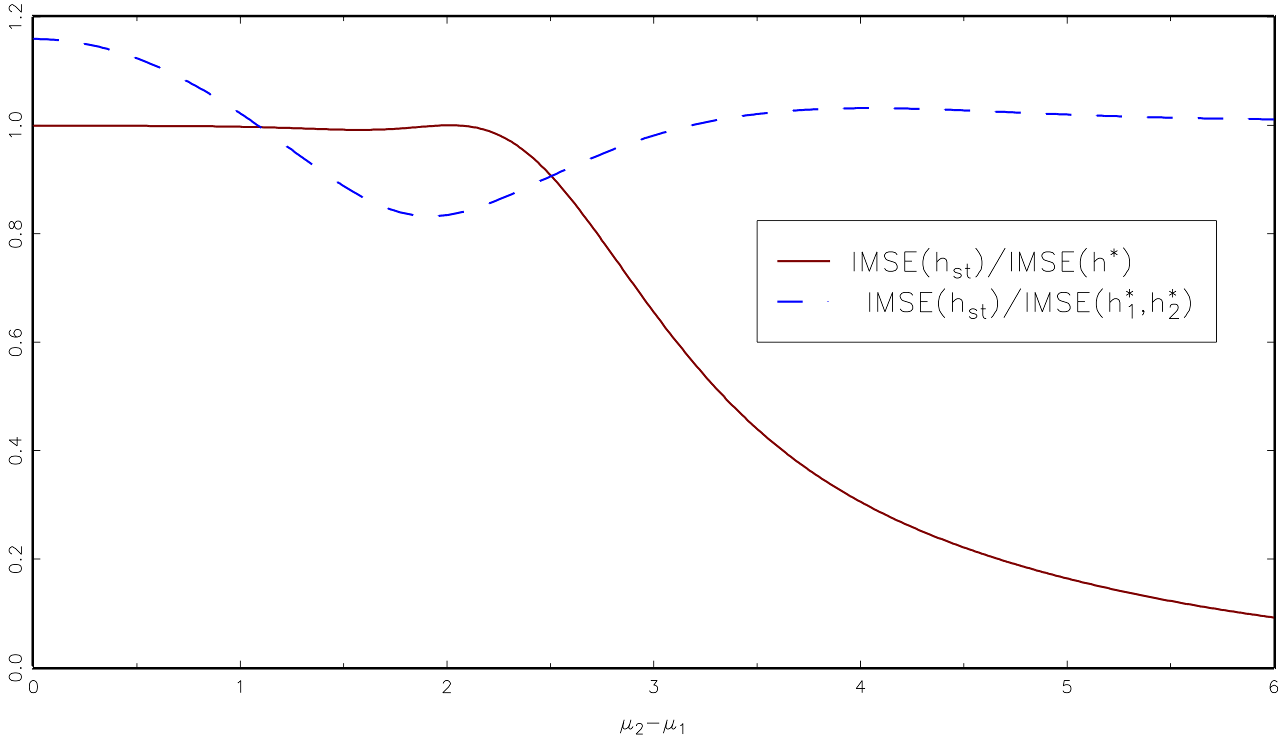


Figure3b

Ratio of Integrated Mean Squared Errors

$\mu_1 = \mu_2 = 0$ $n_1=50, n_2=100$

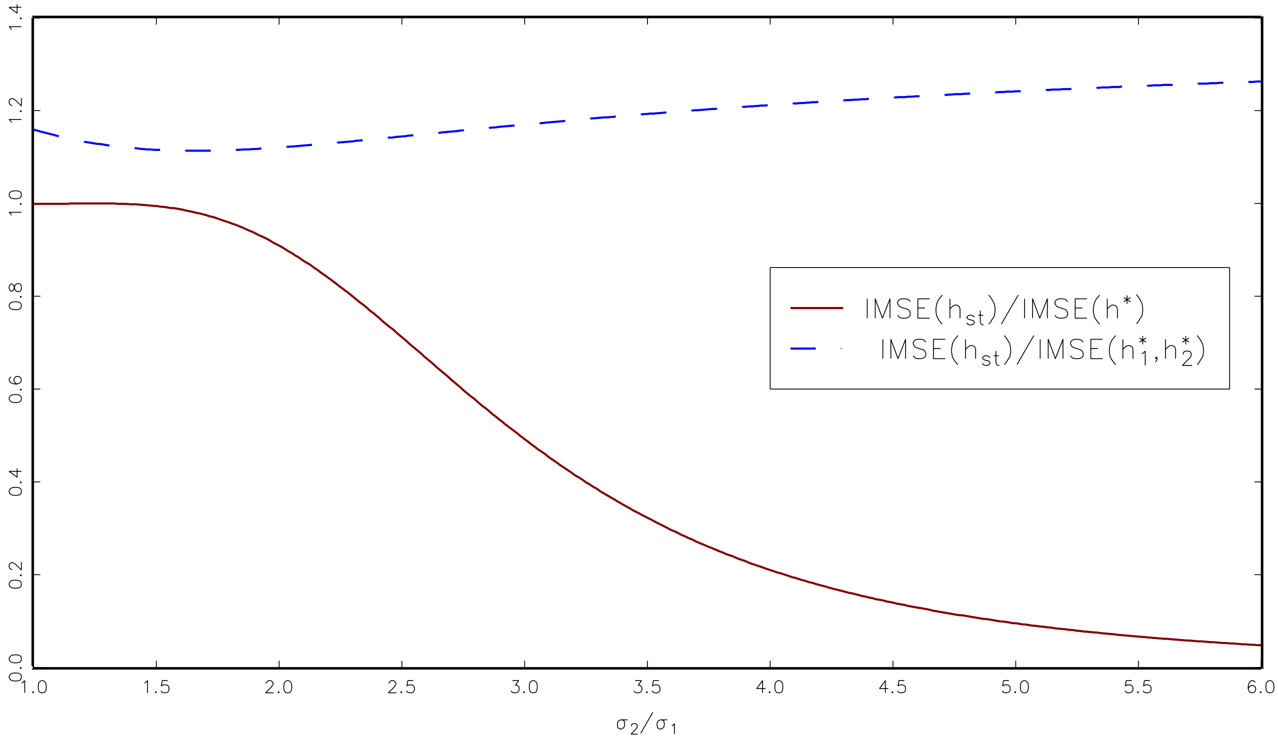


Figure4a
Weighted estimate using h_{st}
Proportional Sampling

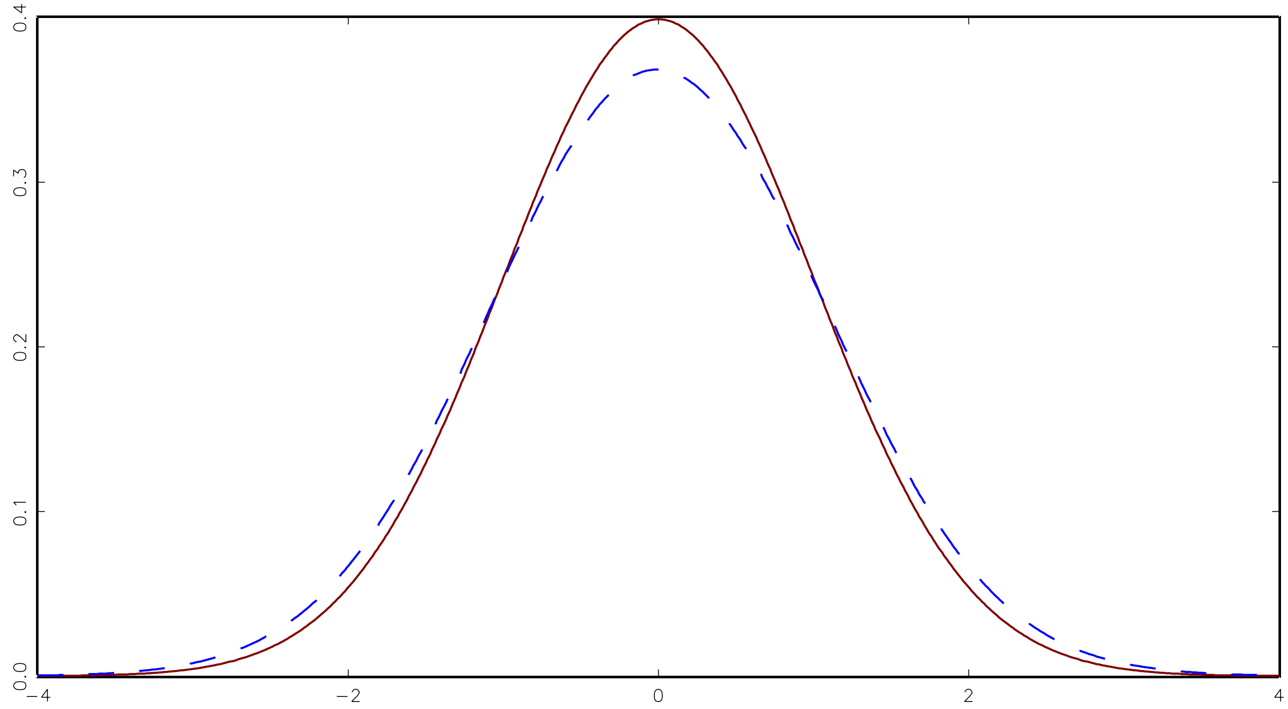


Figure4b
Unweighted estimate using h^*
Proportional Sampling

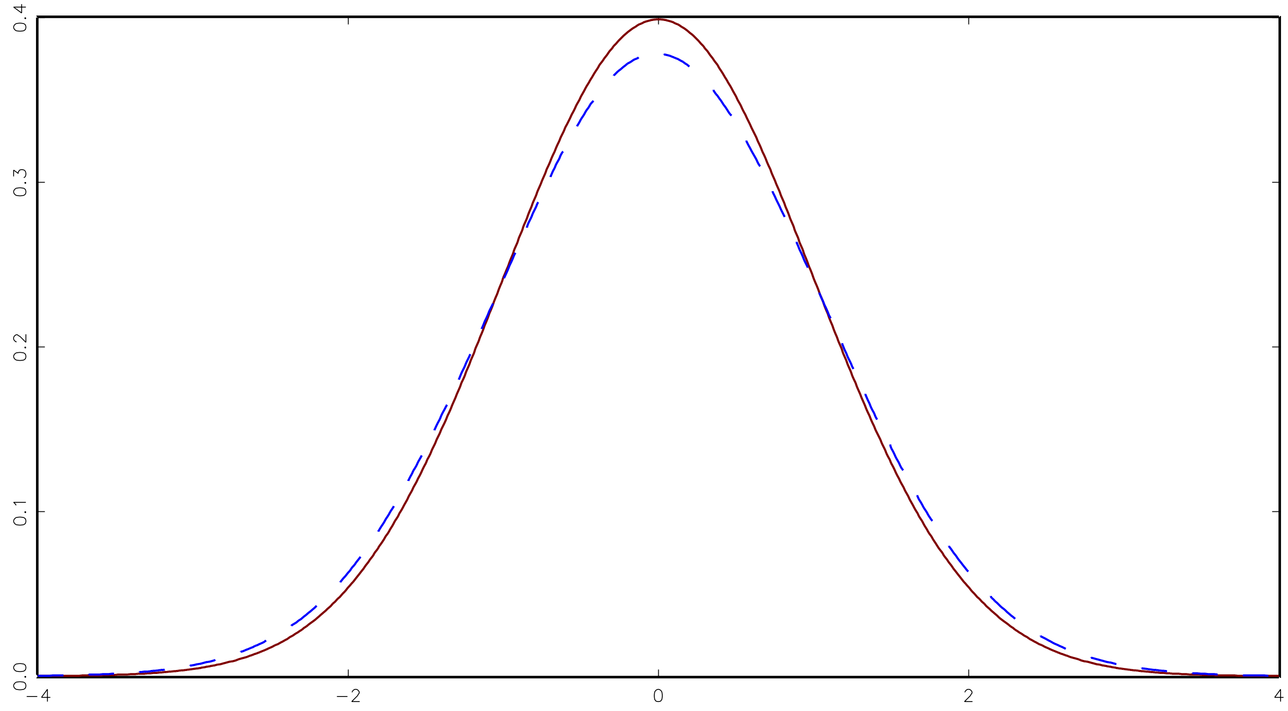


Figure4c
Unweighted estimate using h_a
Proportional Sampling

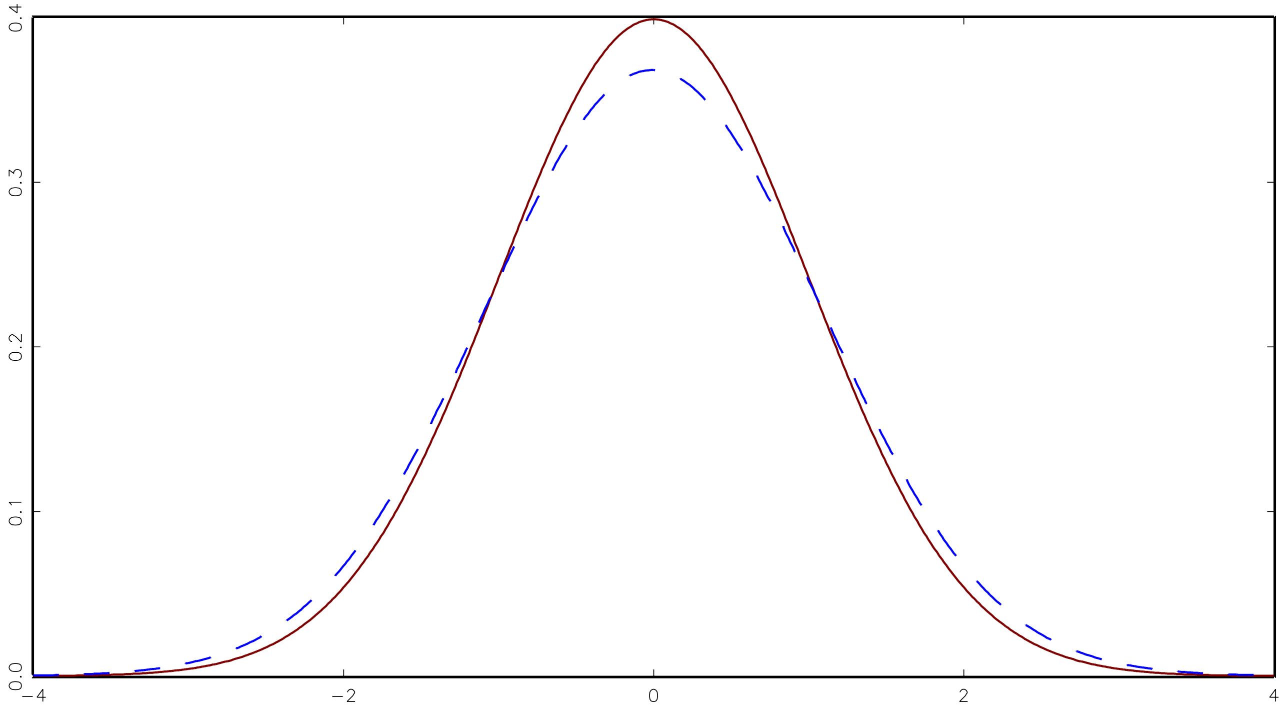
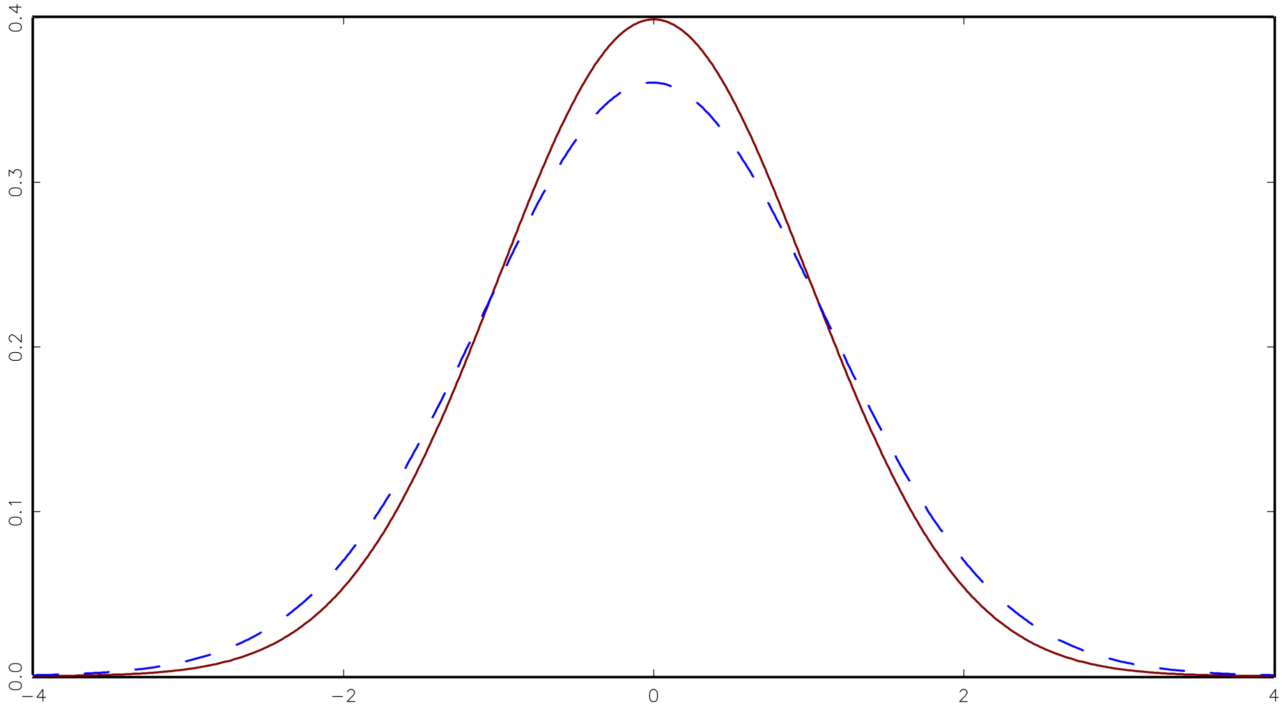


Figure4d
Weighted combination of stratum−specific densitiies
Proportional Sampling



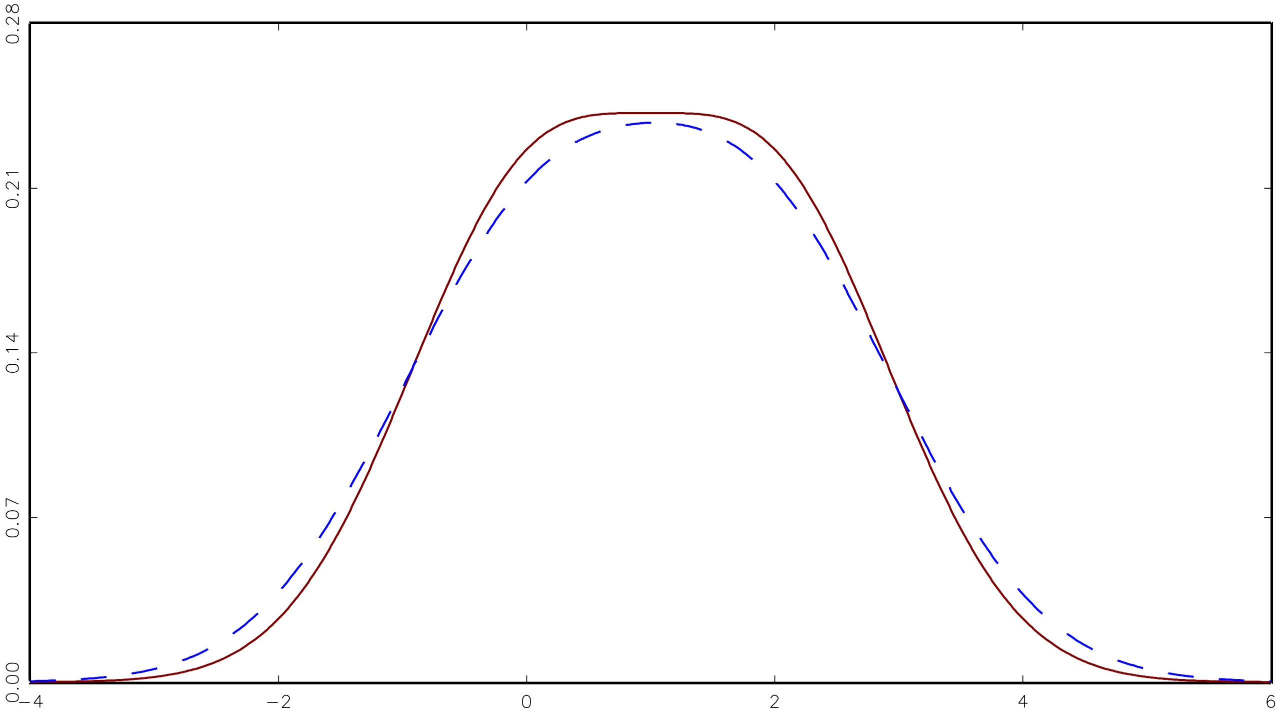


Figure5a
Weighted estimate using h_{st}
Proportional Sampling

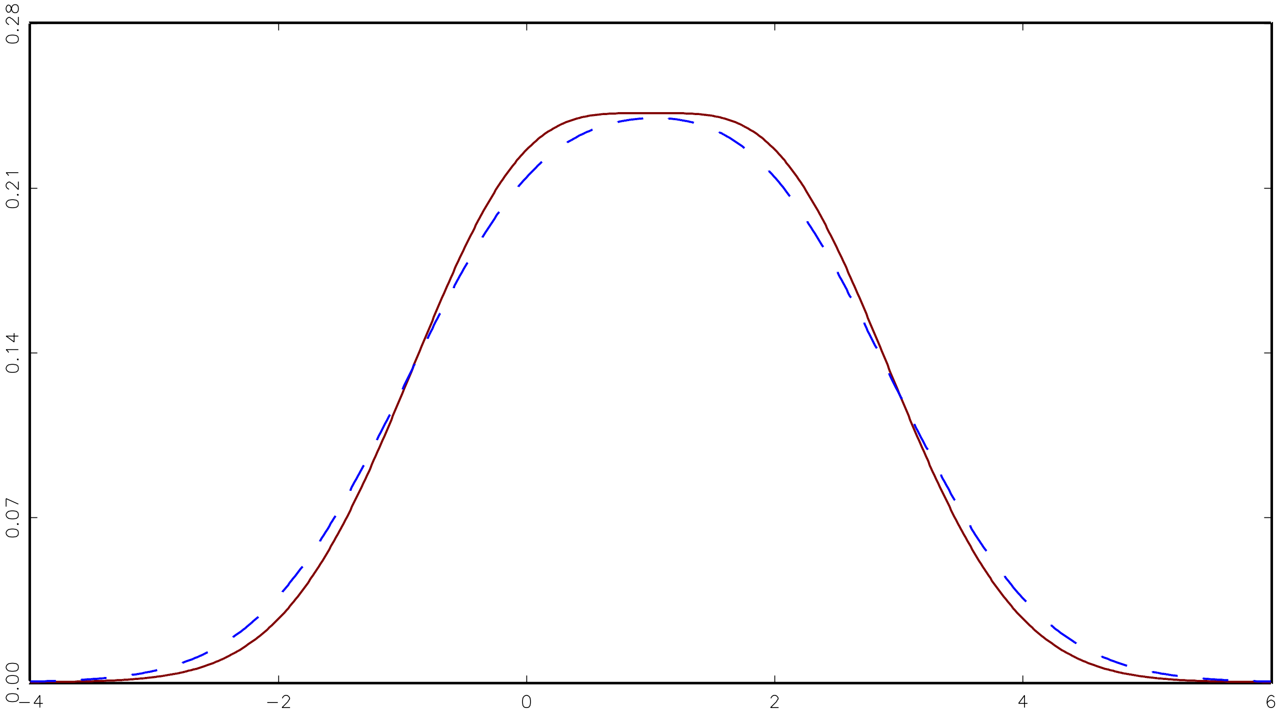


Figure5b
Unweighted estimate using h^*
Proportional Sampling

Figure5c
Unweighted estimate using h_a
Proportional Sampling

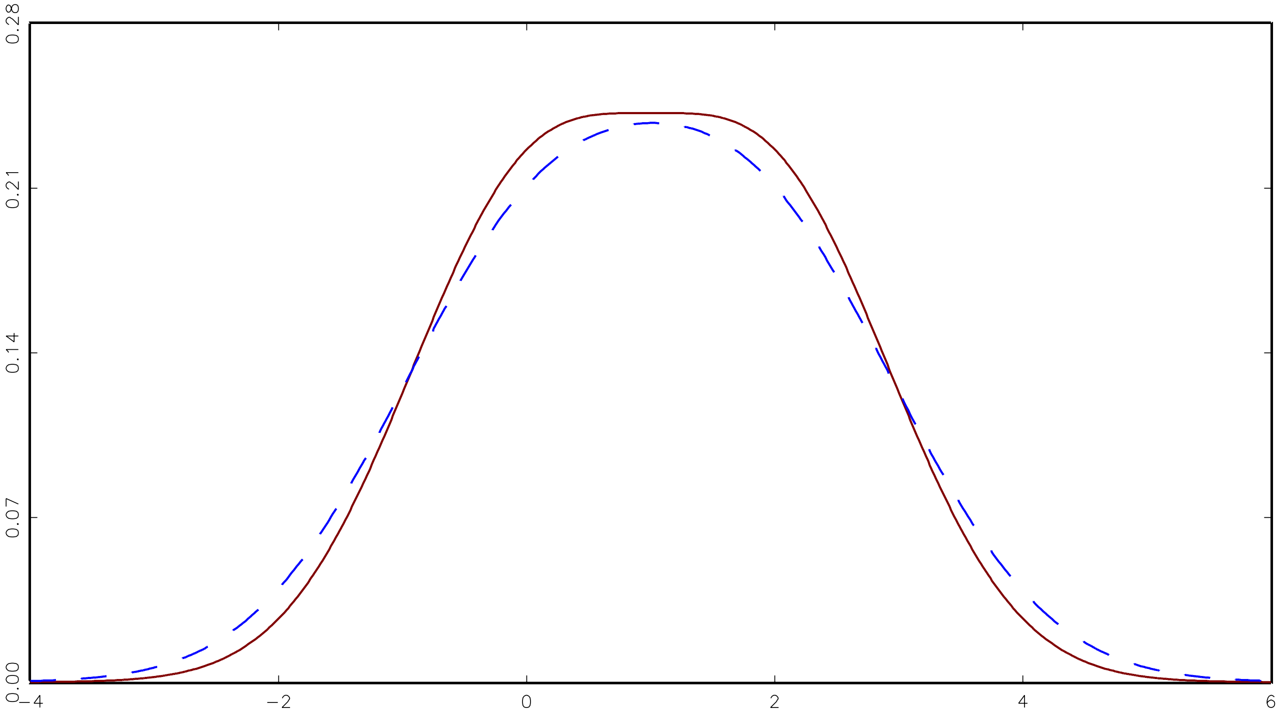


Figure5d
Weighted combination of stratum−specific densitiies
Proportional Sampling

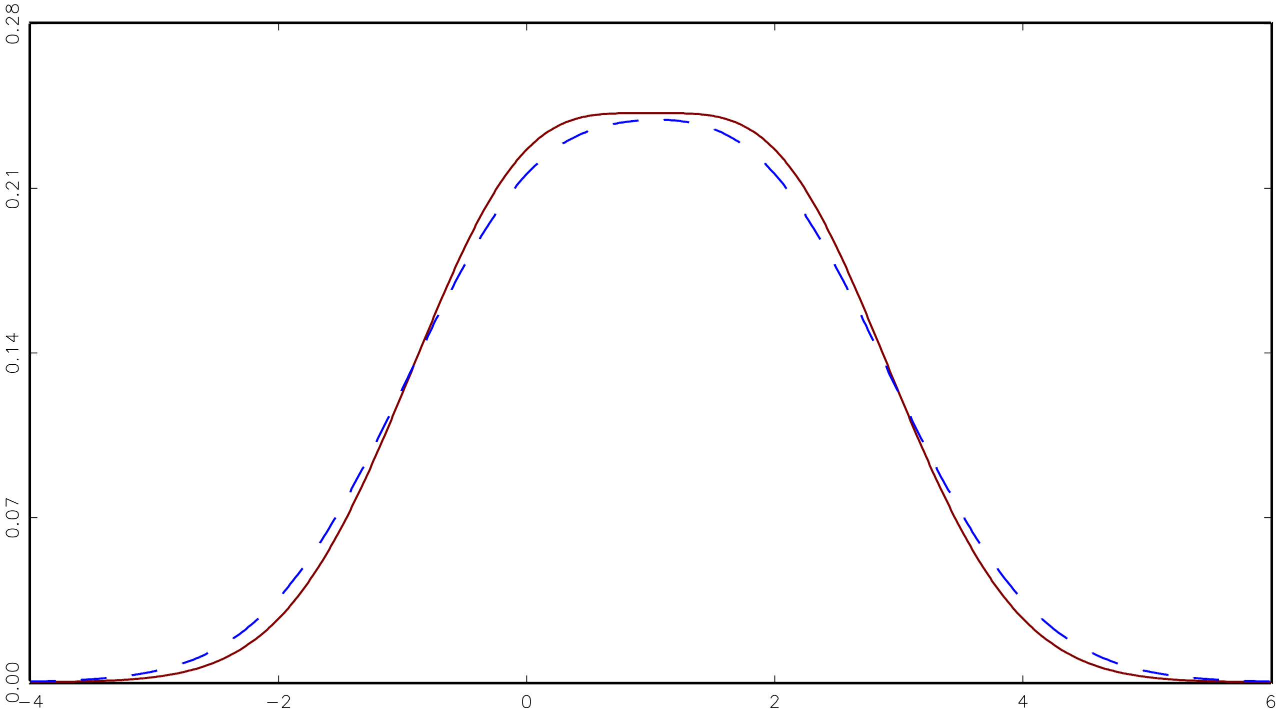


Figure6a
Weighted estimate using h_{st}
Proportional Sampling

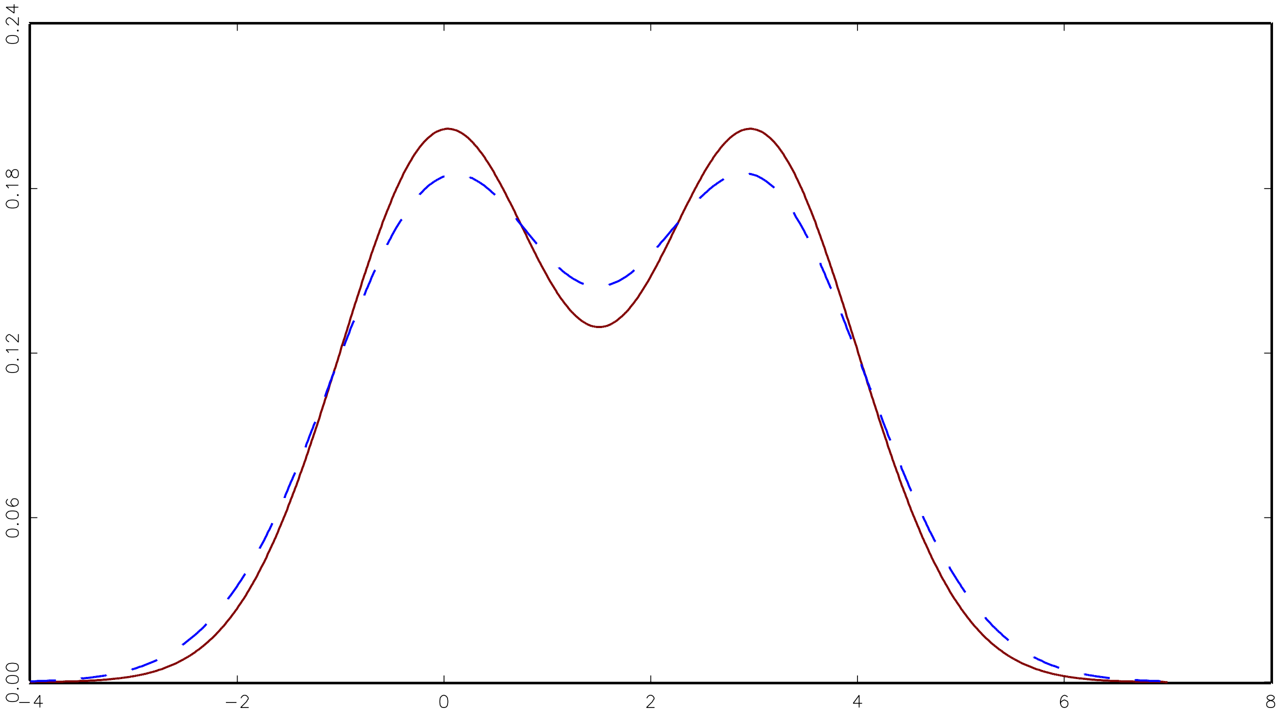


Figure6b
Unweighted estimate using h^*
Proportional Sampling

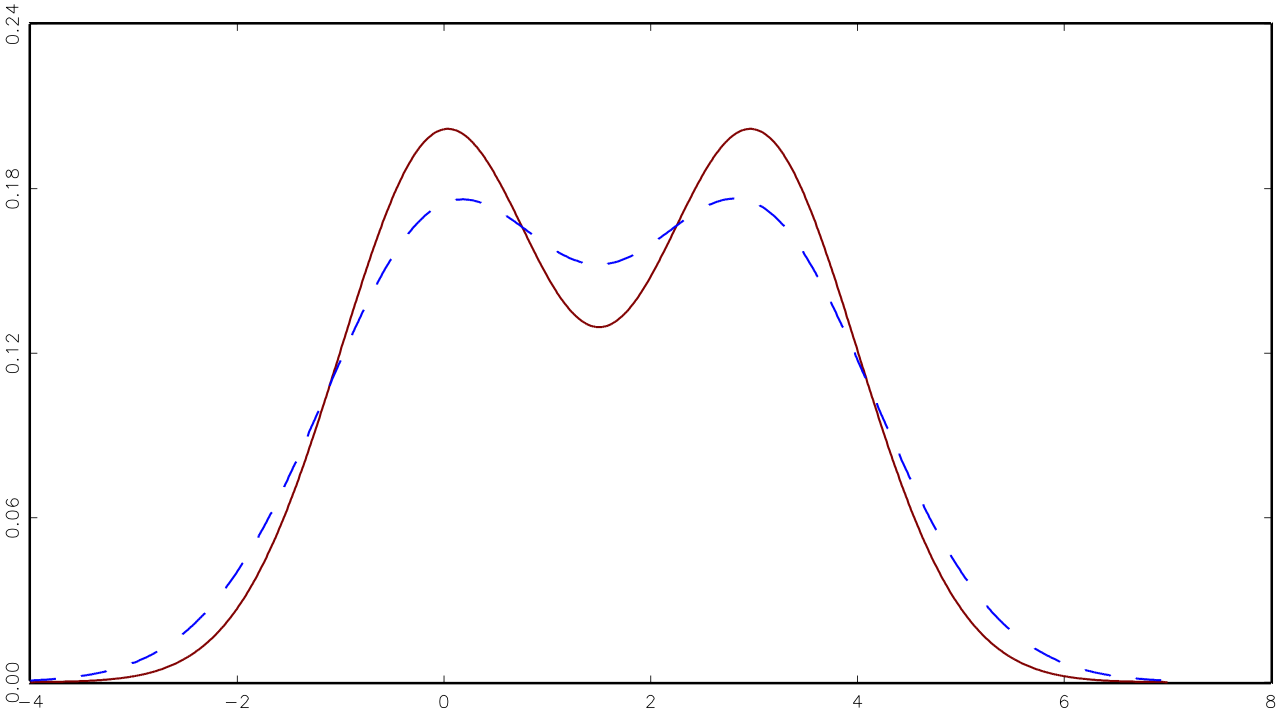


Figure6c
Unweighted estimate using h_a
Proportional Sampling

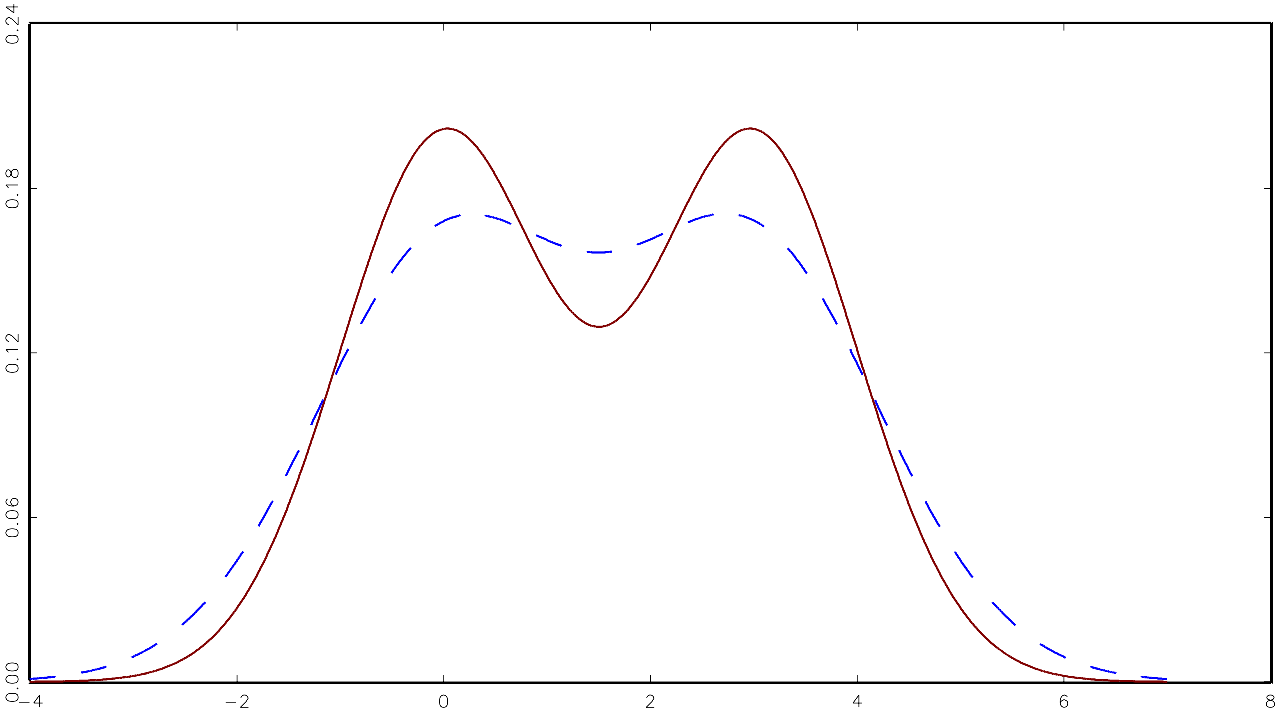


Figure6d
Weighted combination of stratum−specific densitiies
Proportional Sampling

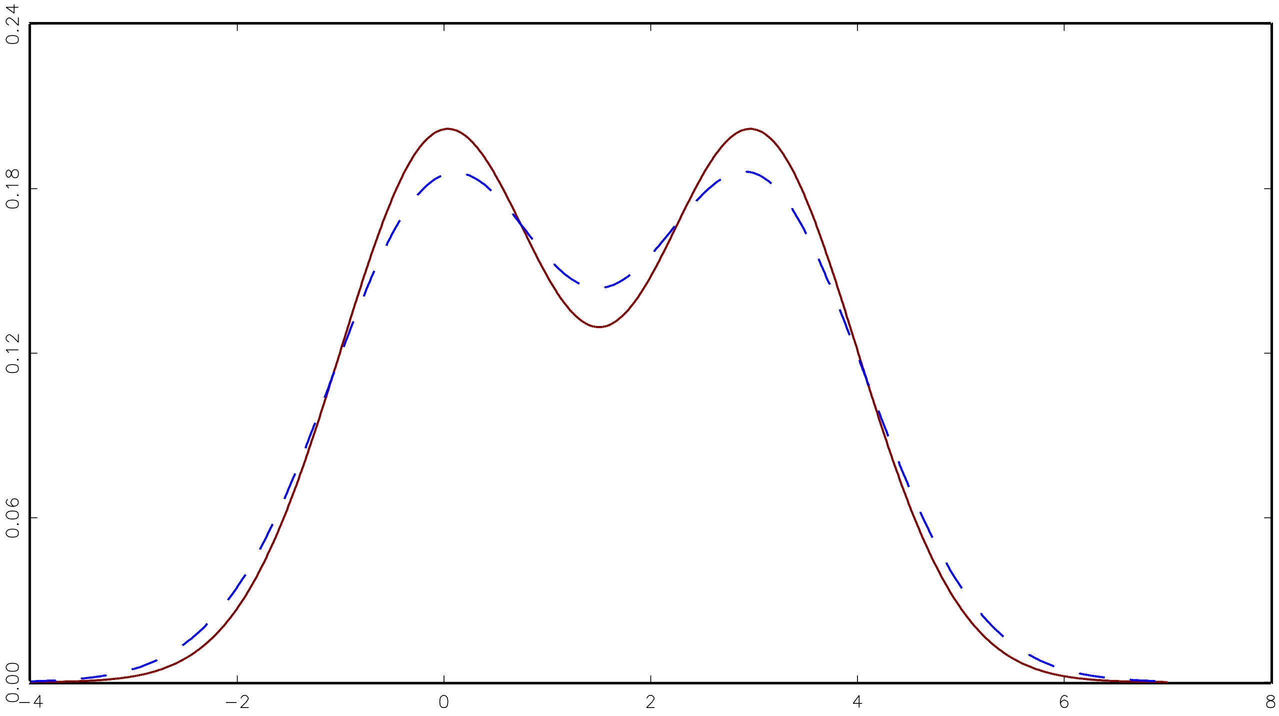


Figure7a

Weighted estimate using h_{st}

Proportional Sampling

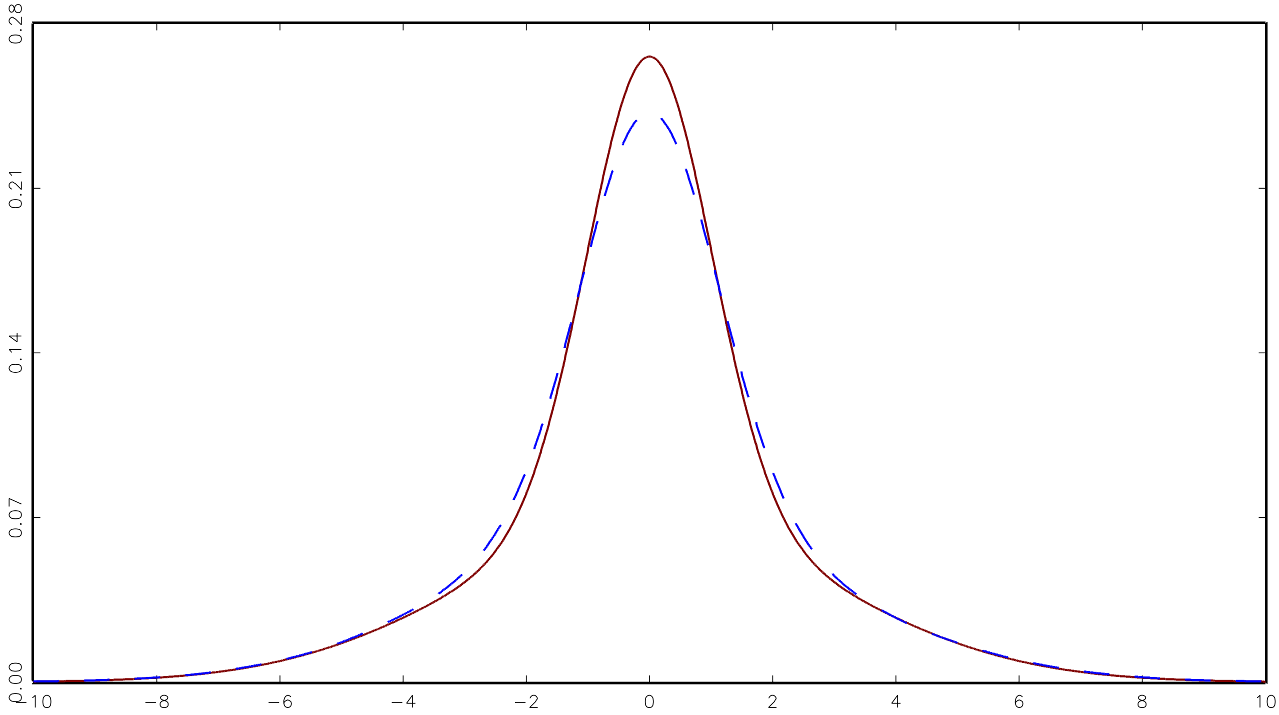


Figure7b

Unweighted estimate using h^*

Proportional Sampling

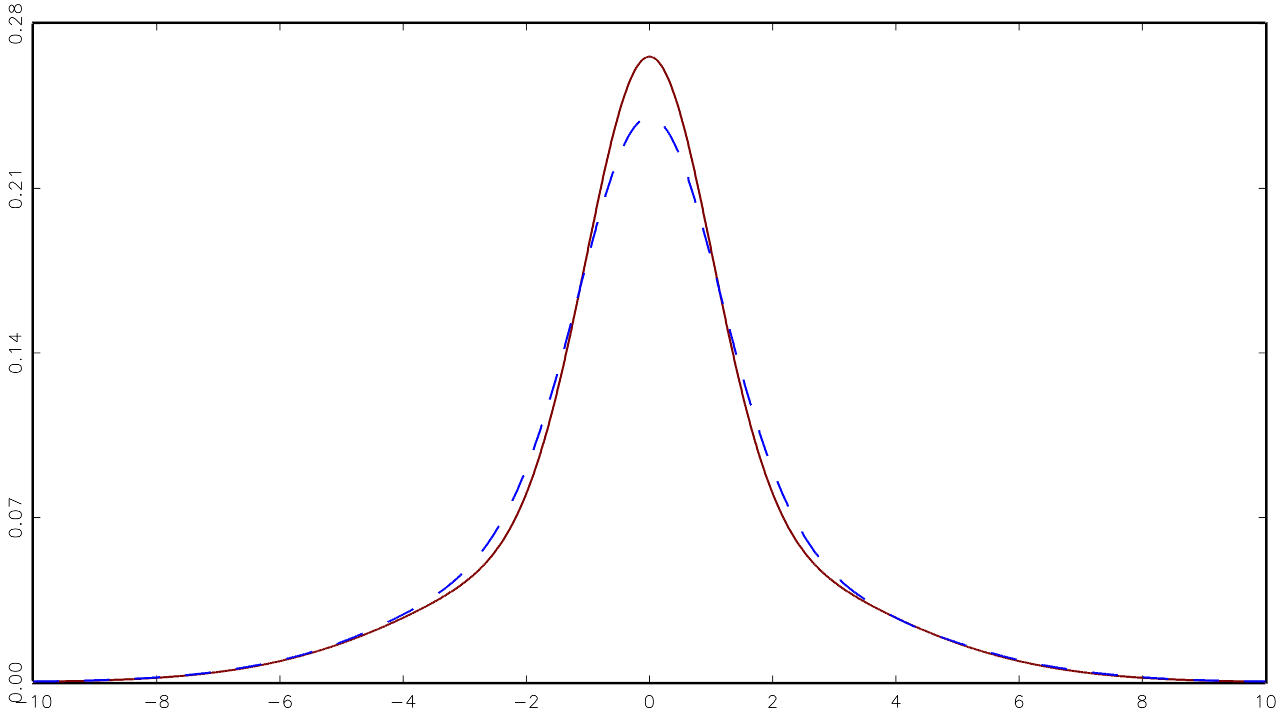


Figure7c
Unweighted estimate using h_a
Proportional Sampling

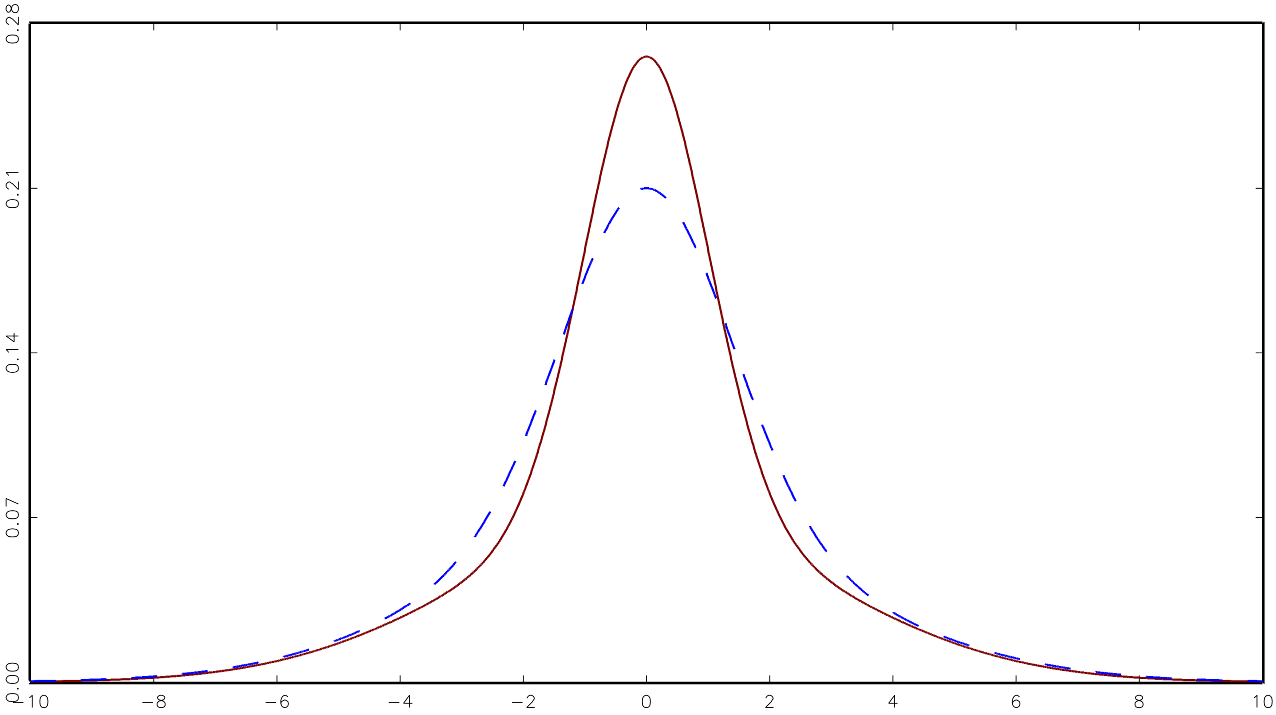


Figure7d
Weighted combination of stratum−specific densitiies
Proportional Sampling

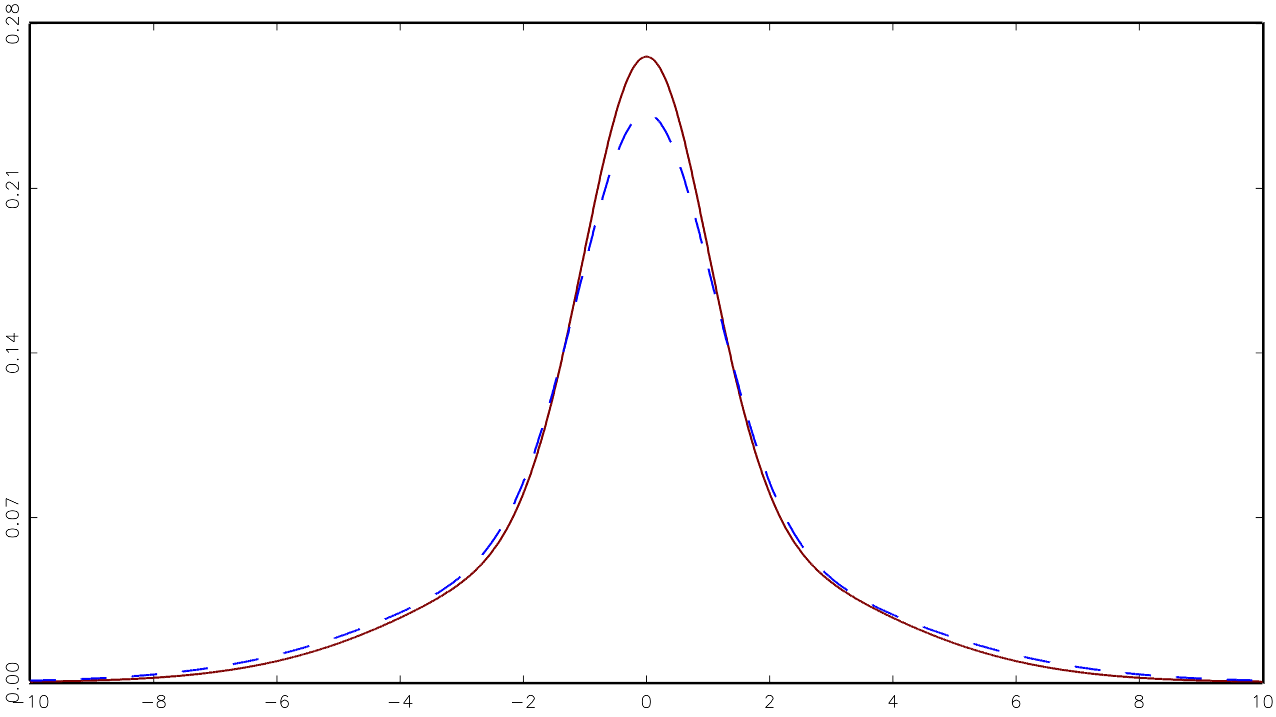


Figure8a
Weighted estimate using h_{st}
Proportional Sampling

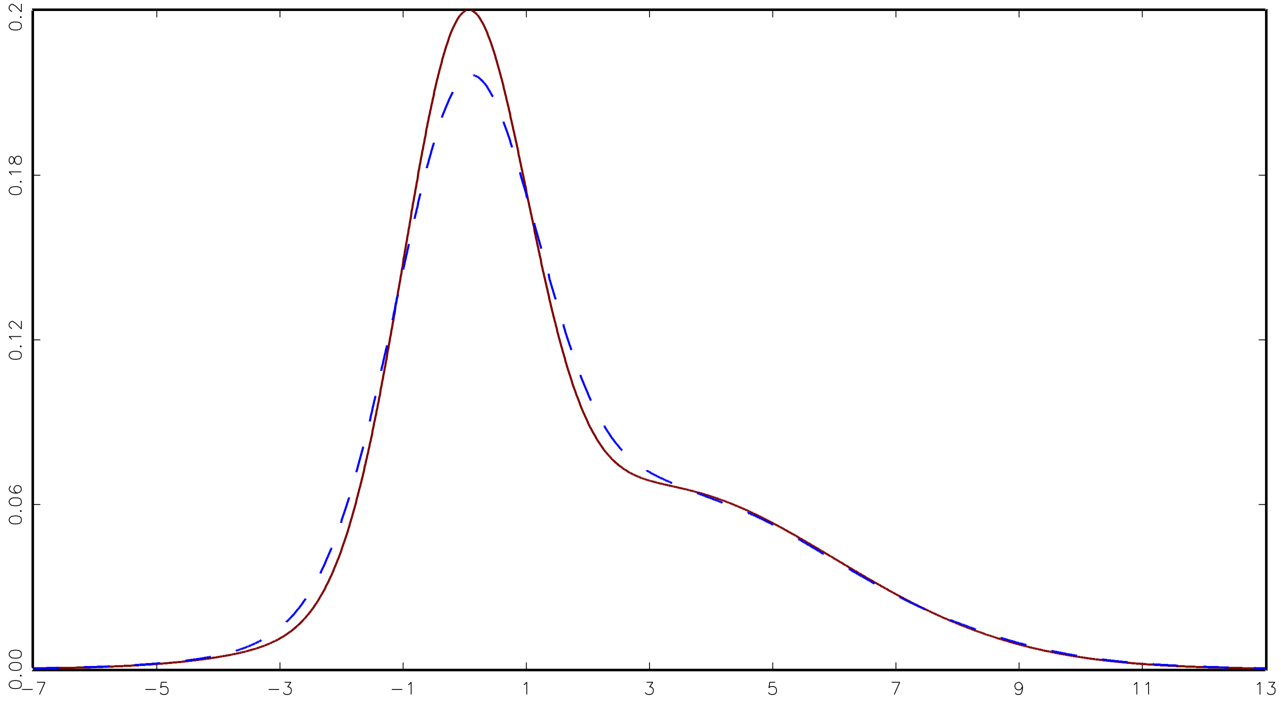


Figure8b
Unweighted estimate using h^*
Proportional Sampling

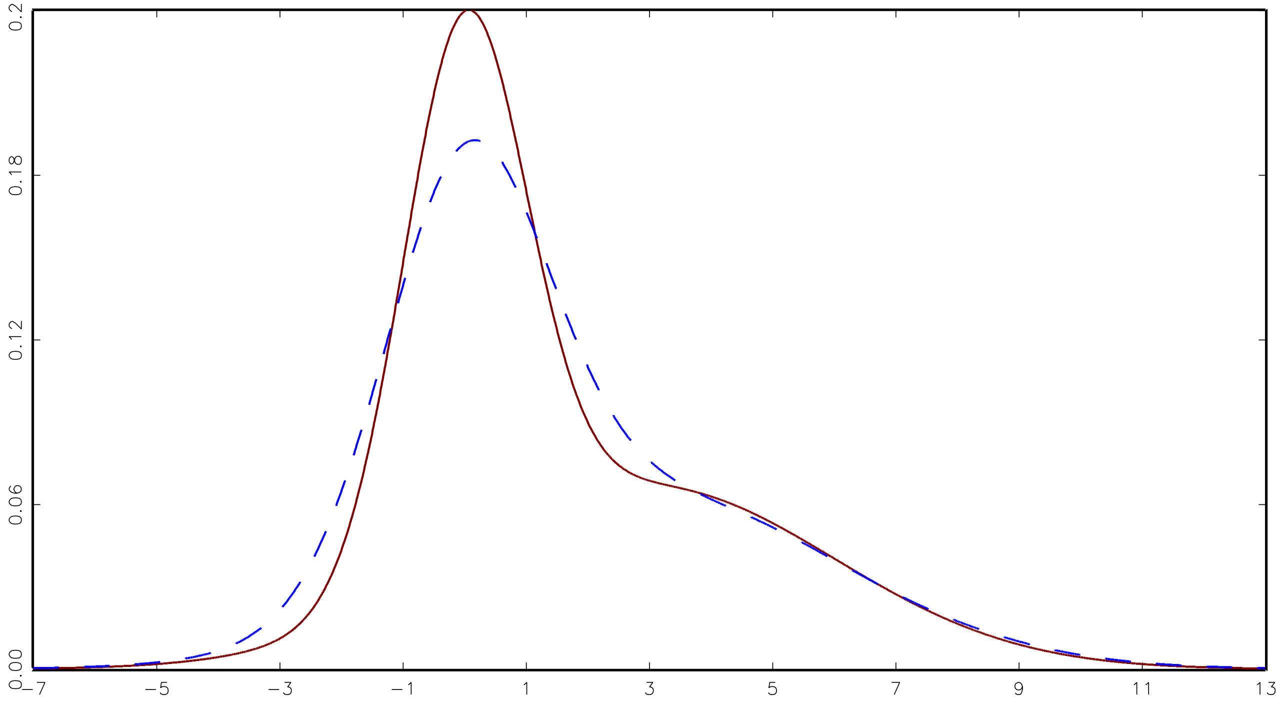


Figure8c
Unweighted estimate using h_a
Proportional Sampling

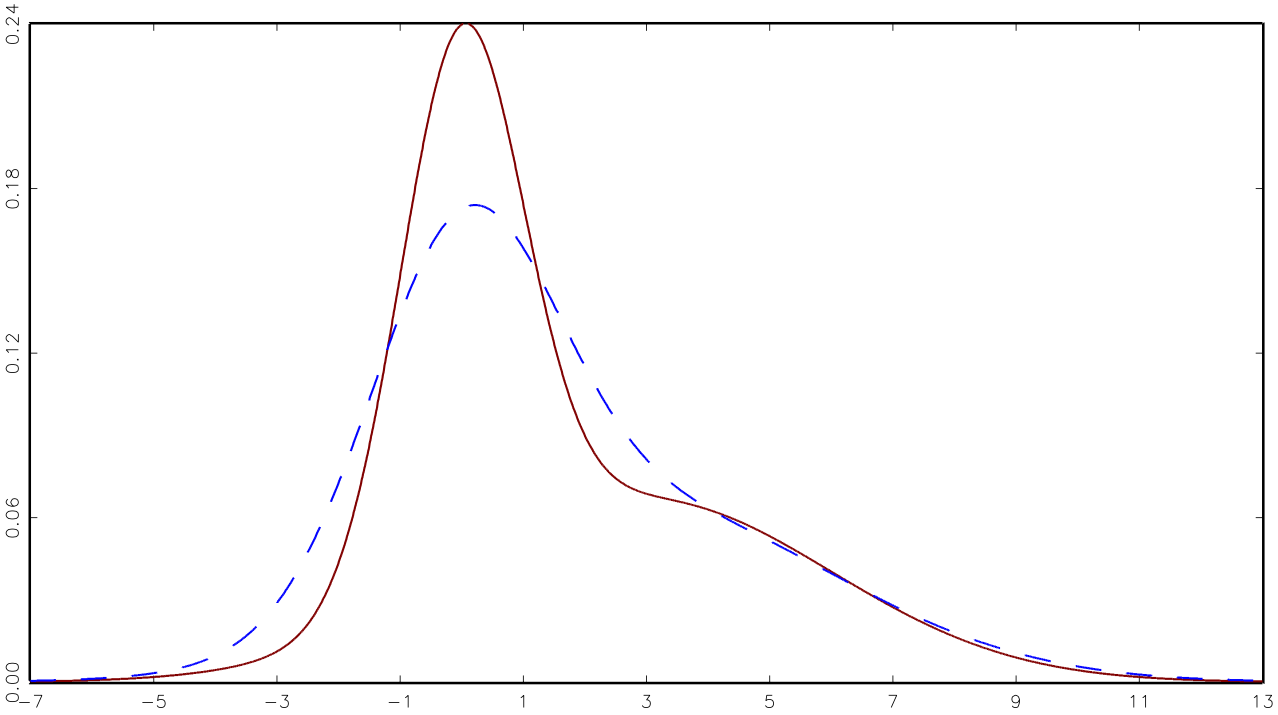


Figure8d
Weighted combination of stratum−specific densitiies
Proportional Sampling

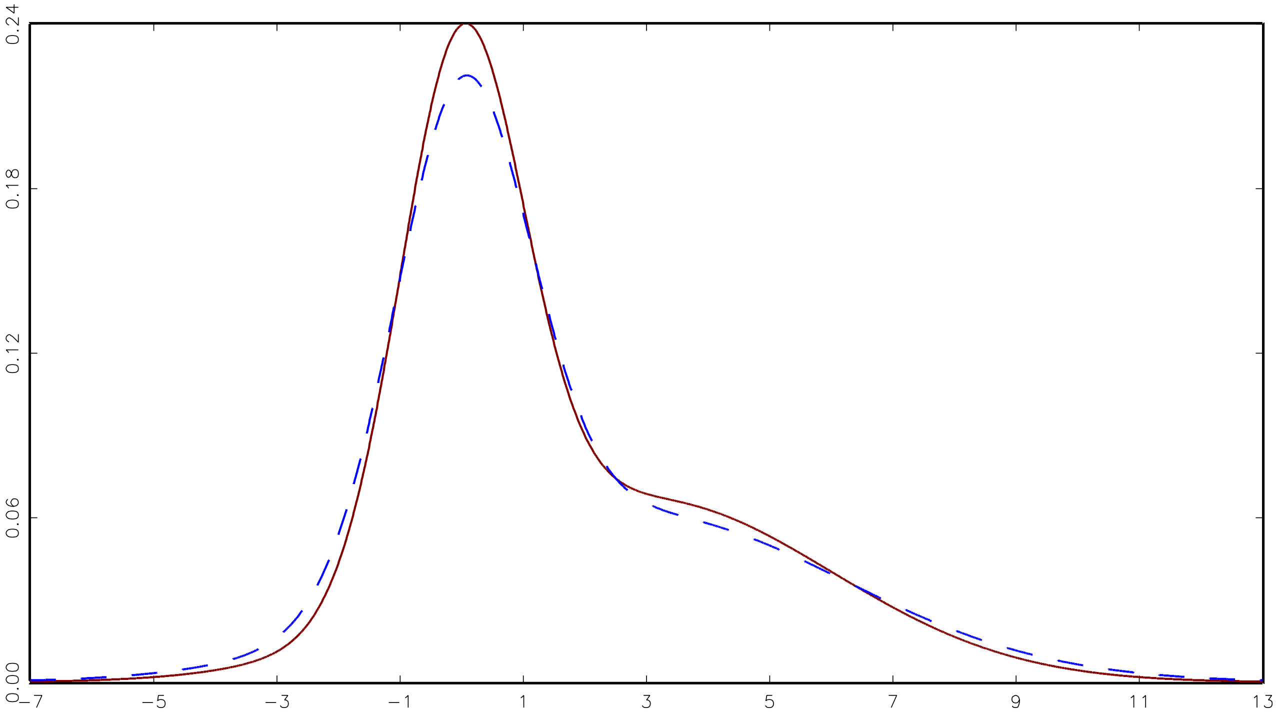


Figure9a

Weighted estimate using h_{st}

Non−proportional Sampling n2/n1=2

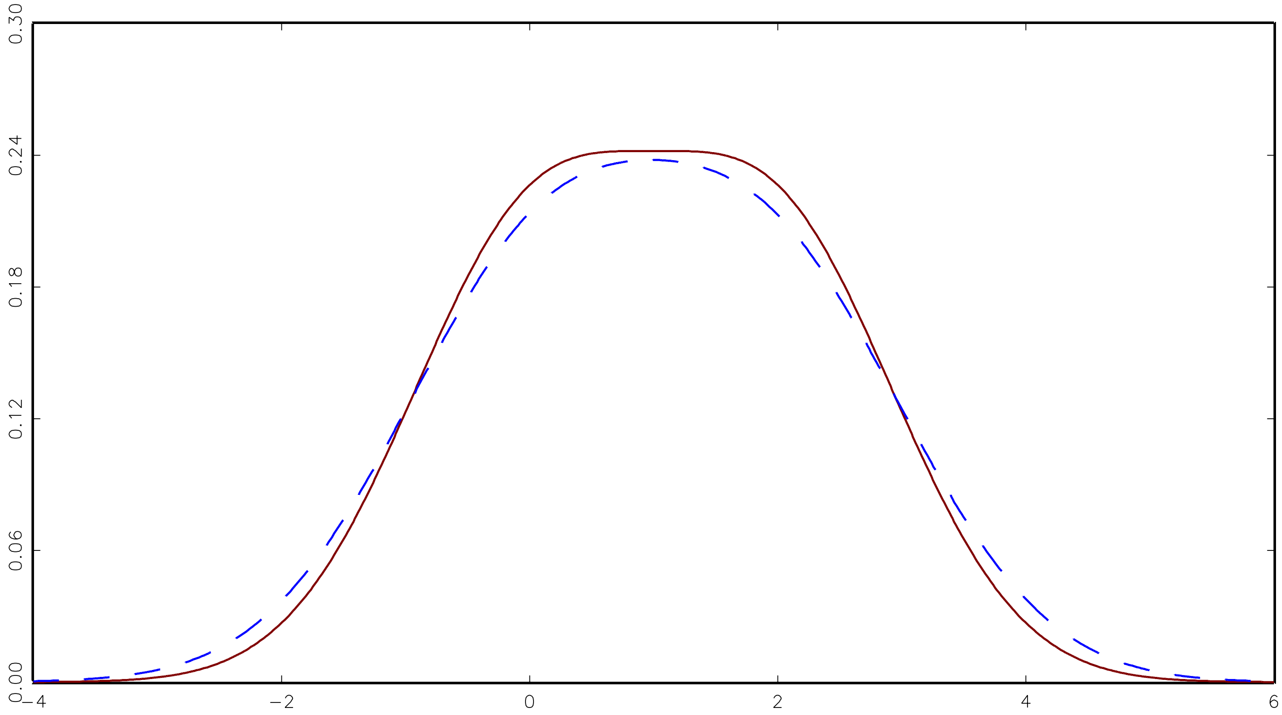


Figure9b

Unweighted estimate using h^*

Non−proportional Sampling n2/n1=2

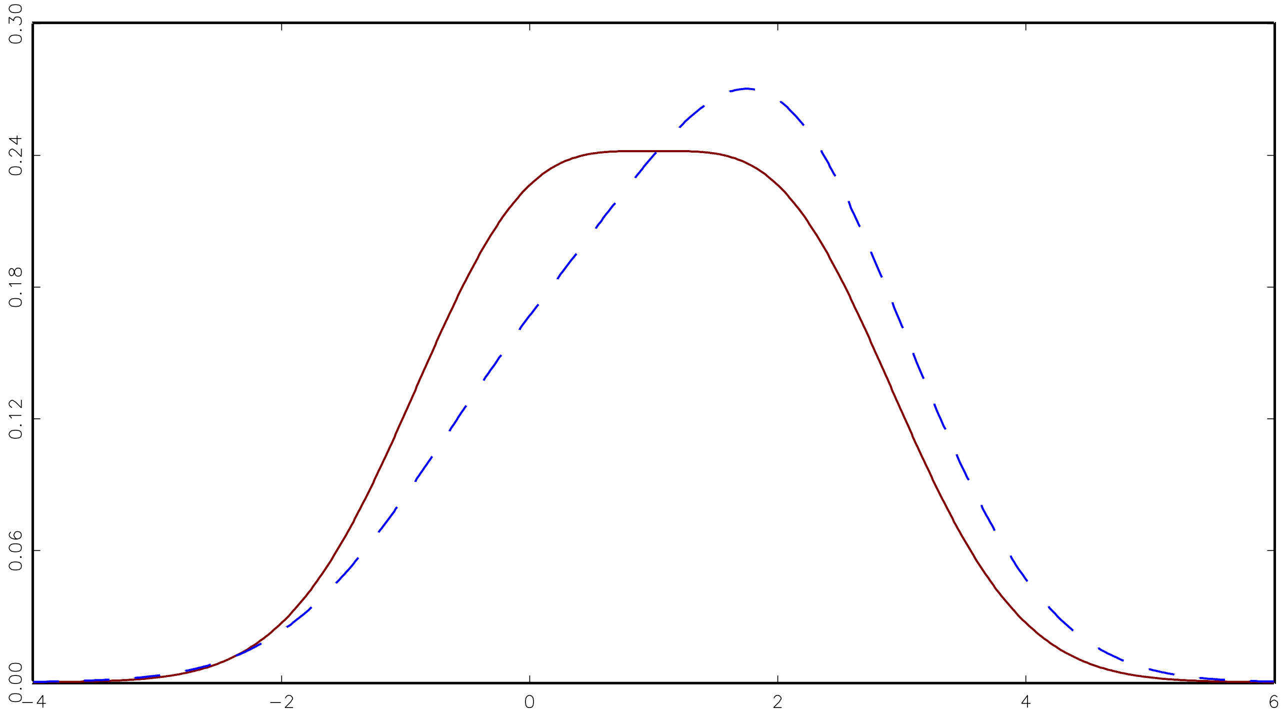


Figure9c
Unweighted estimate using h_a
Non−proportional Sampling   n2/n1=2

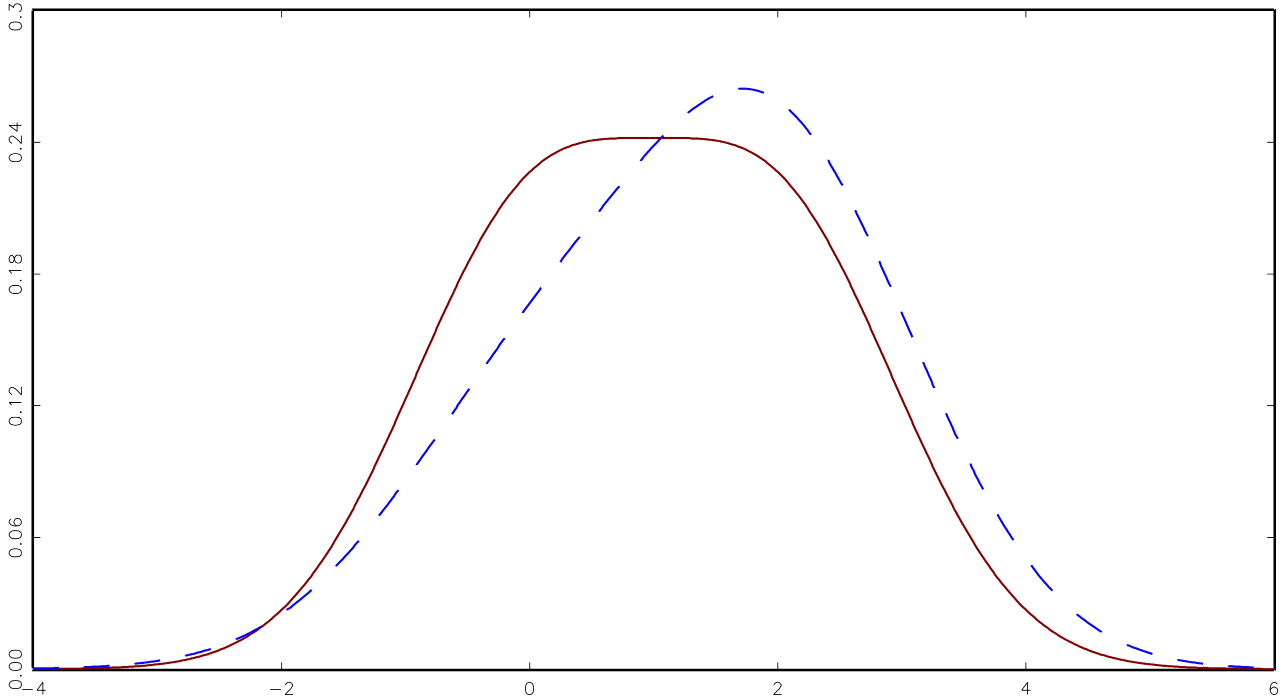
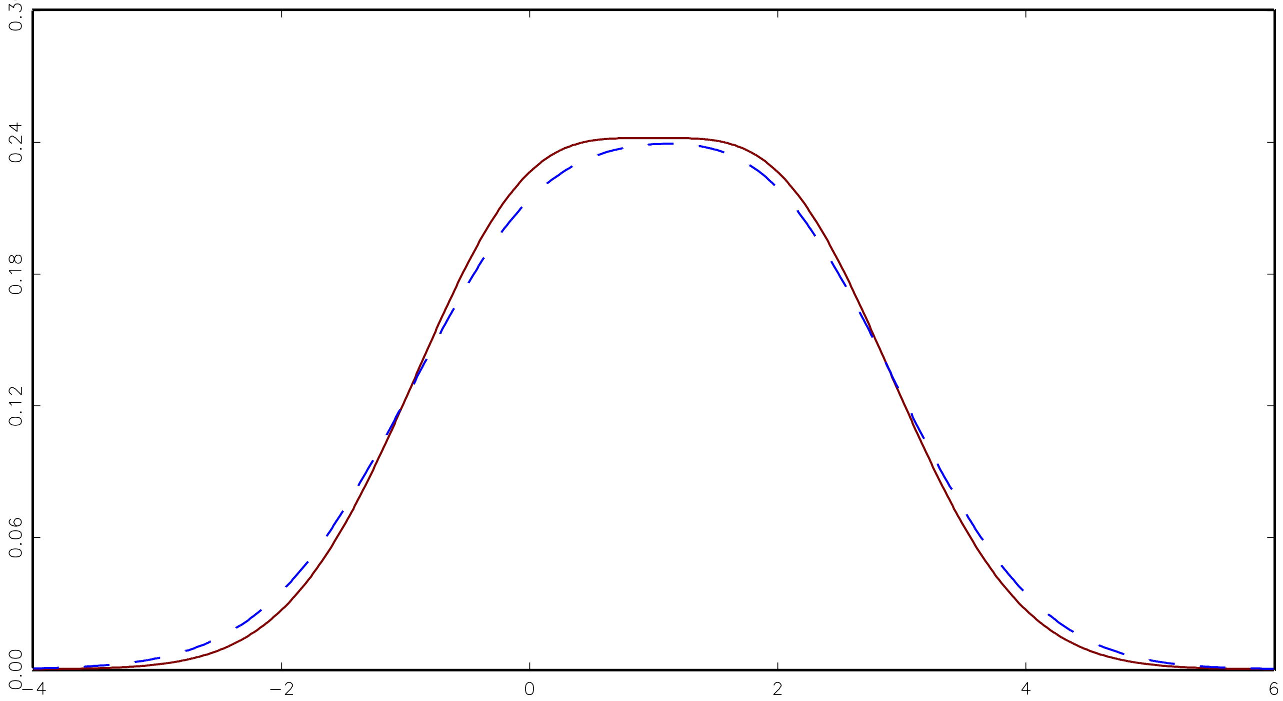


Figure9d
Weighted combination of stratum−specific densitiies
Non−proportional Sampling   n2/n1=2



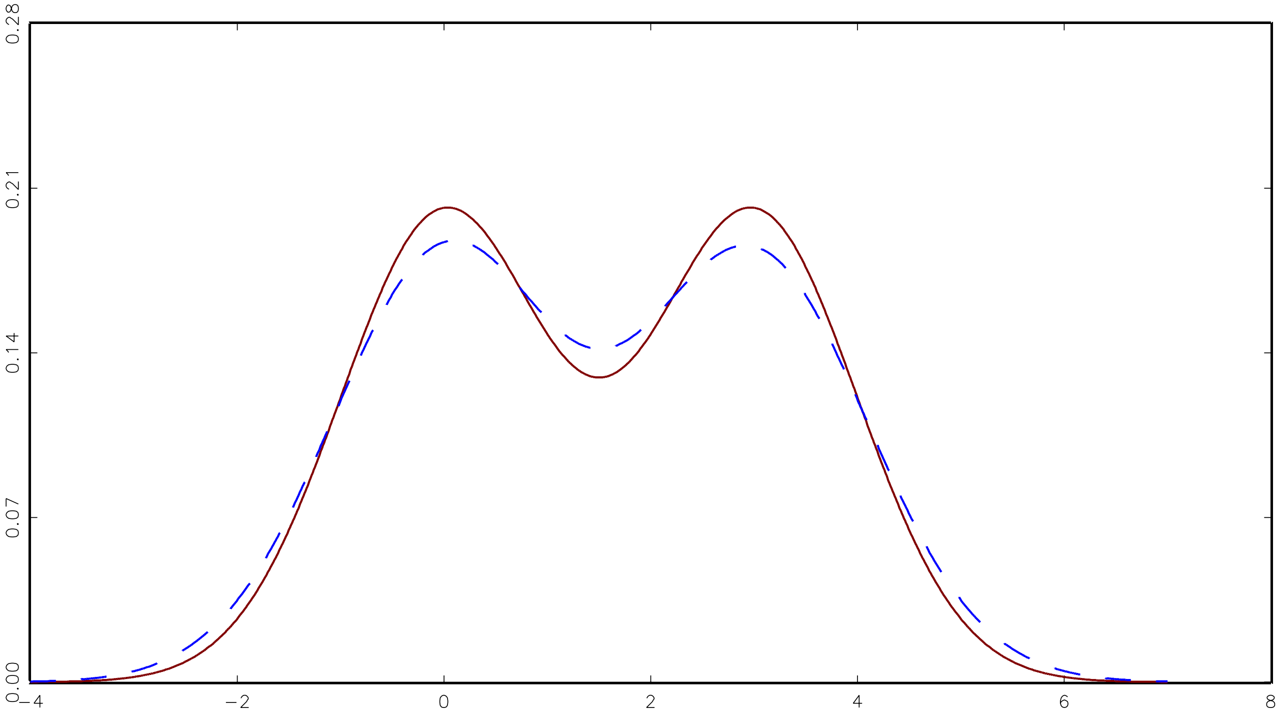


Figure10a
Weighted estimate using h_{st}
Non−proportional Sampling n2/n1=2

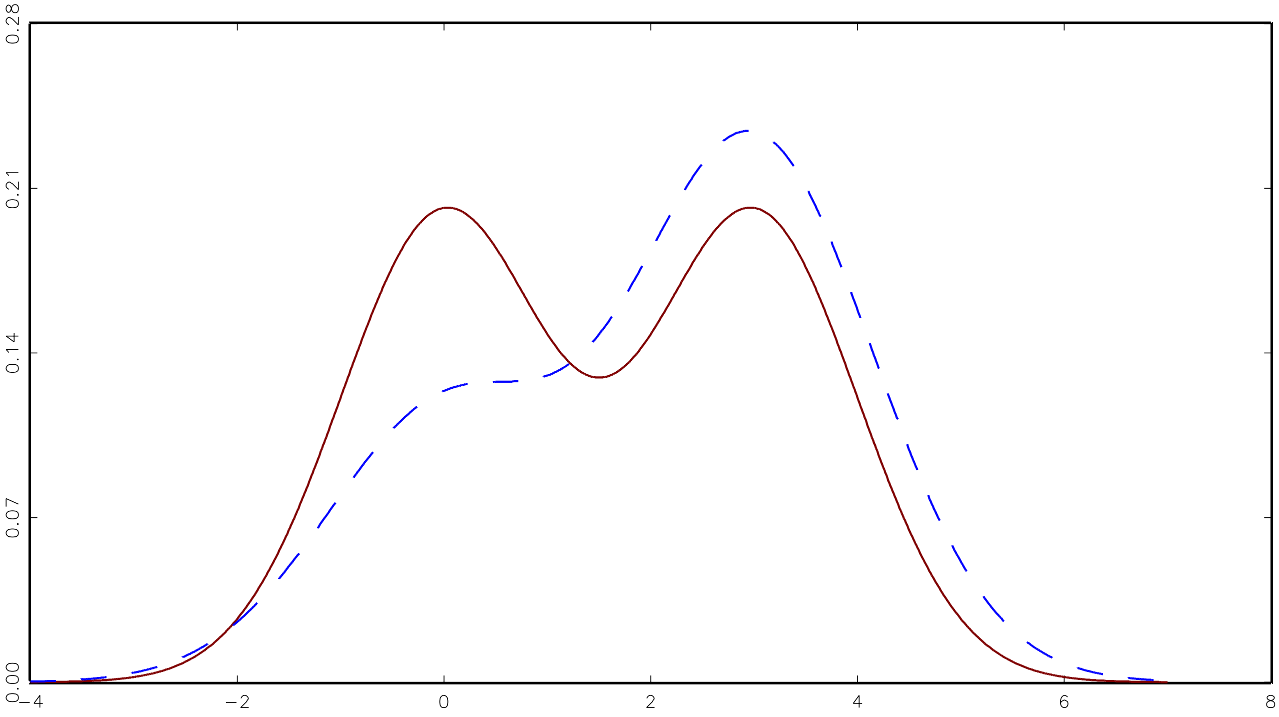


Figure10b
Unweighted estimate using h^*
Non−proportional Sampling n2/n1=2

Figure10c

Unweighted estimate using h_a

Non−proportional Sampling   n2/n1=2

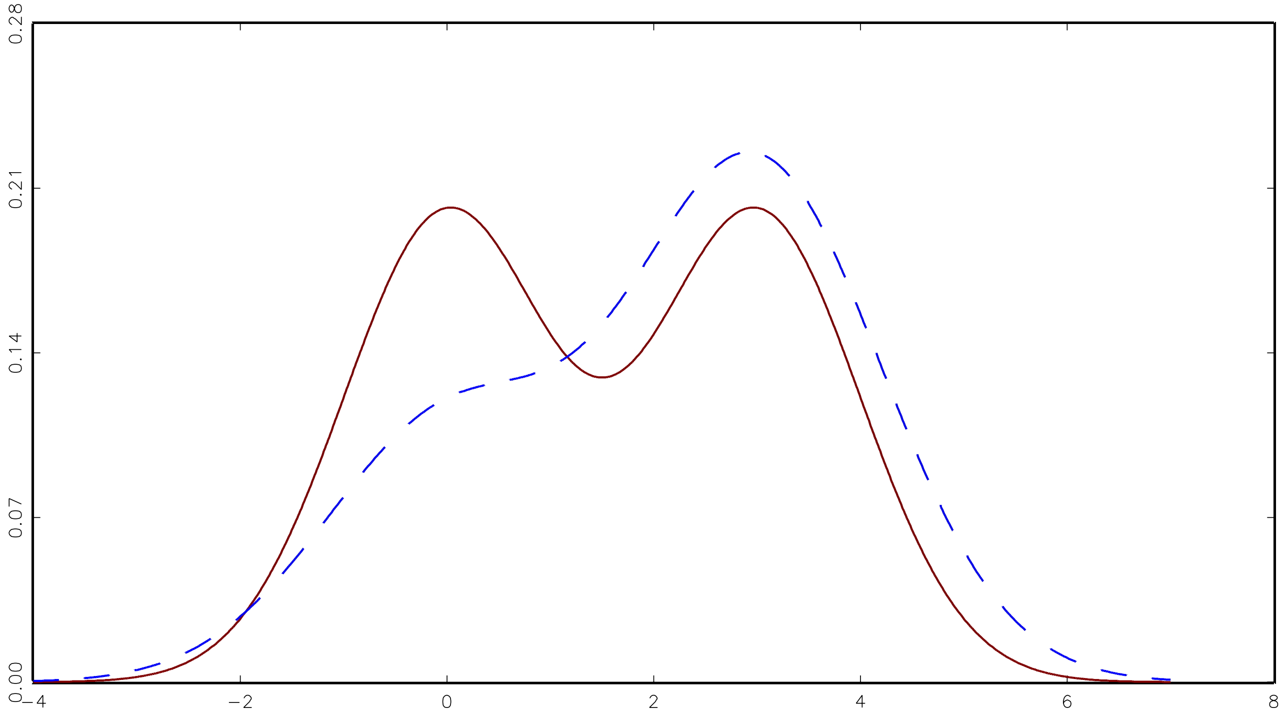


Figure10d

Weighted combination of stratum−specific densitiies

Non−proportional Sampling   n2/n1=2

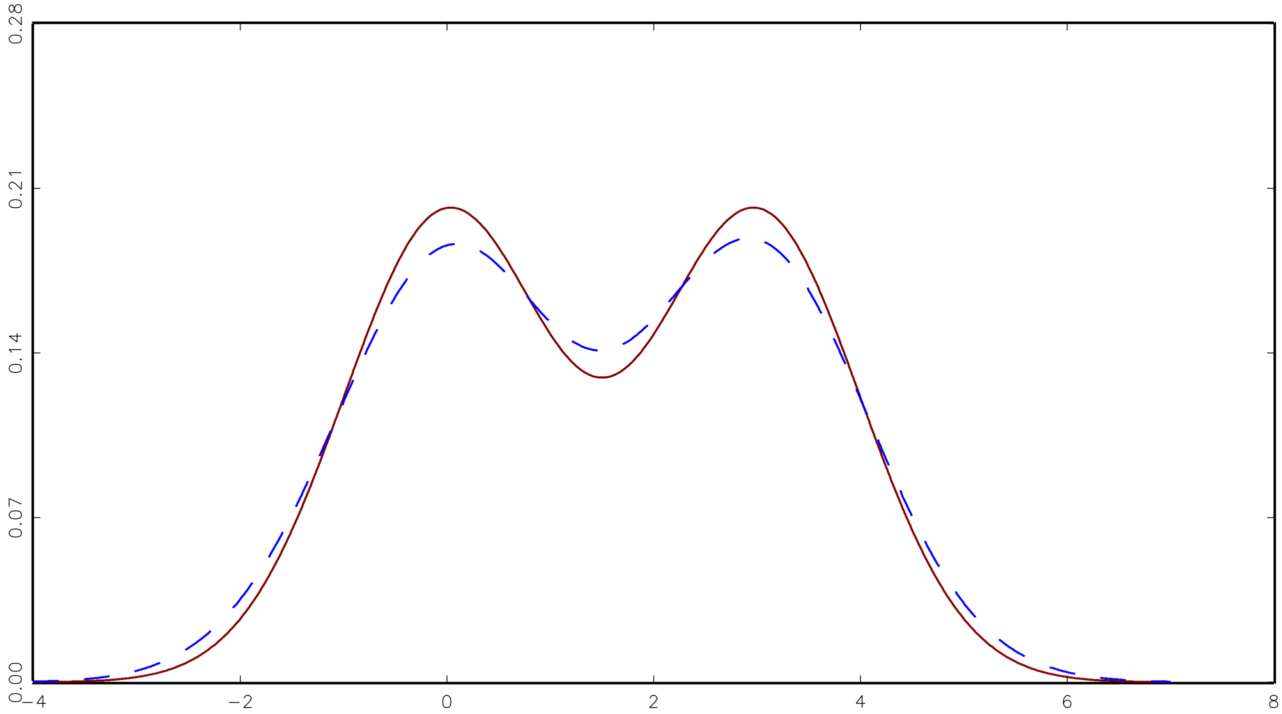


Figure11a

Weighted estimate using h_{st}

Non−proportional Sampling   n2/n1=2

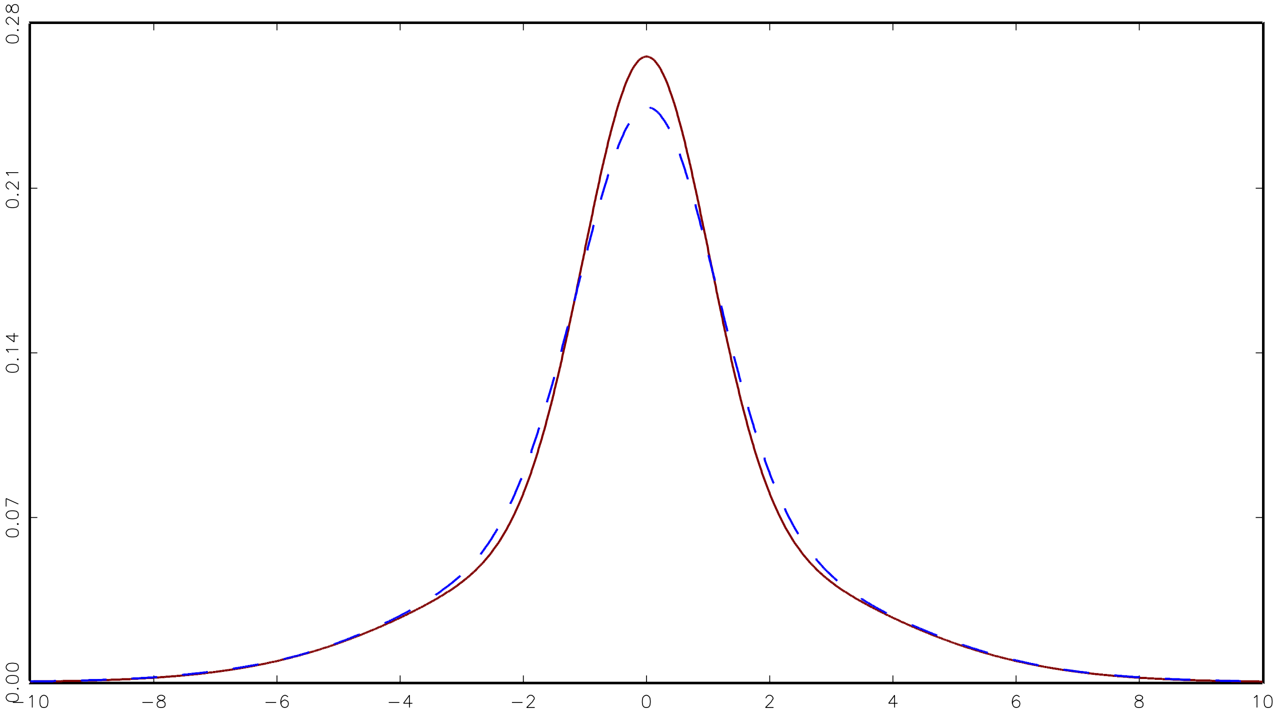


Figure11b

Unweighted estimate using h^*

Non−proportional Sampling   n2/n1=2

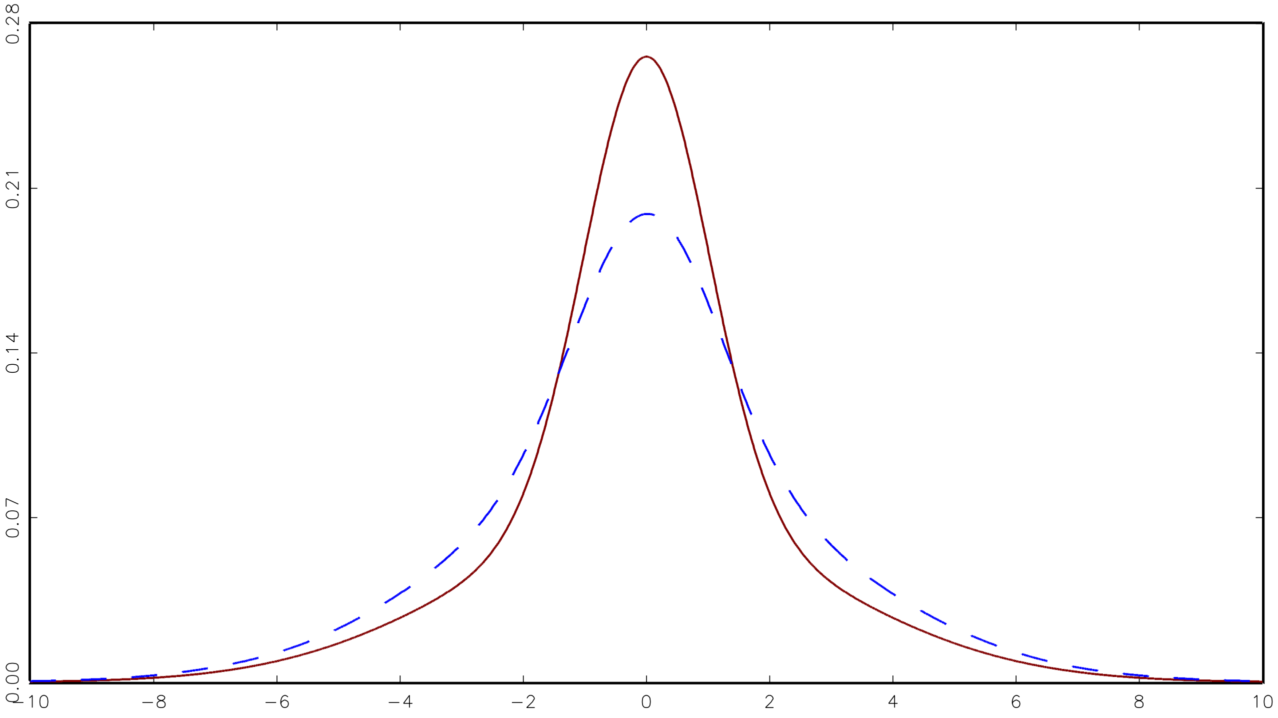


Figure11c
Unweighted estimate using h_a
Non−proportional Sampling   n2/n1=2

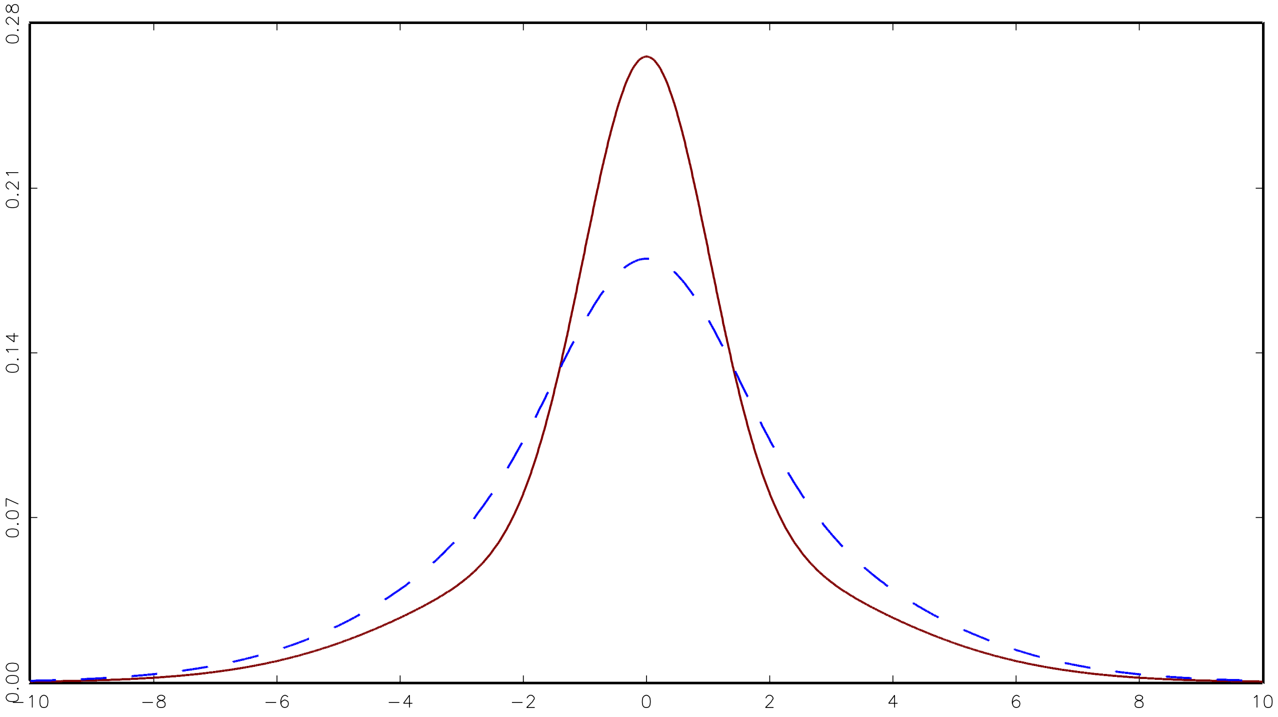


Figure11d
Weighted combination of stratum−specific densitiies
Non−proportional Sampling   n2/n1=2

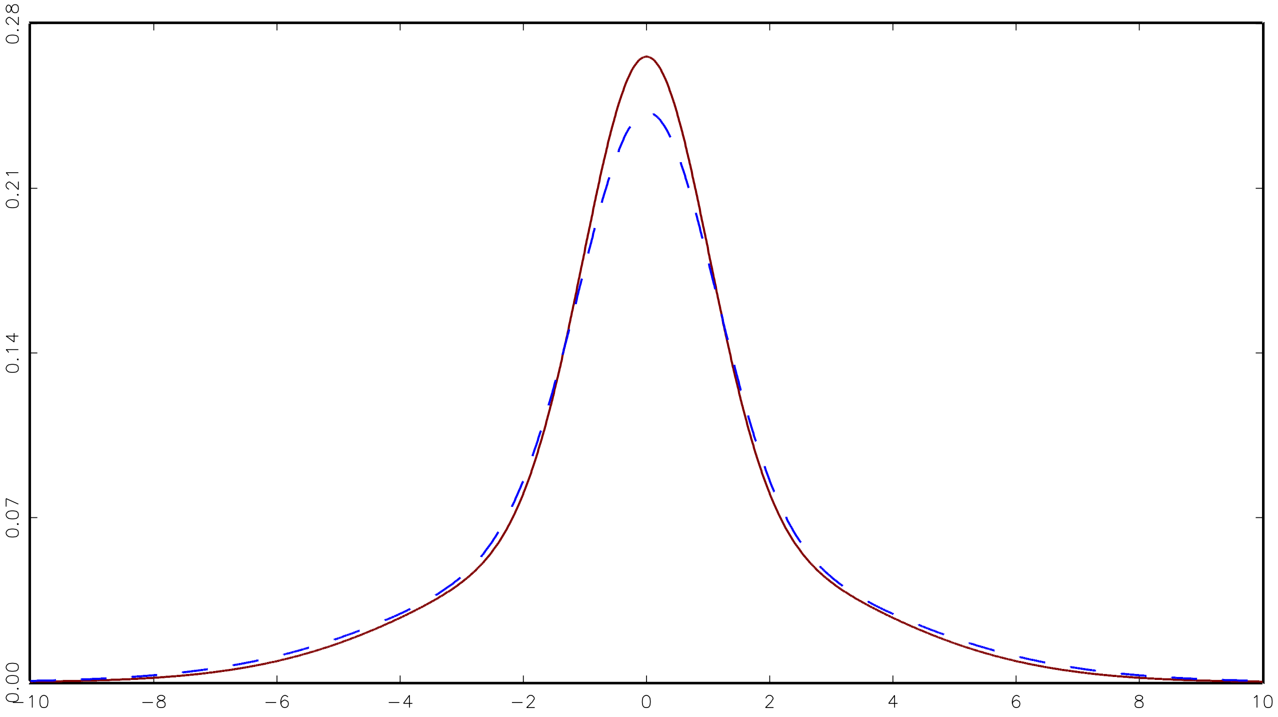


Figure12a

Weighted estimate using h_{st}

Non−proportional Sampling n2/n1=2

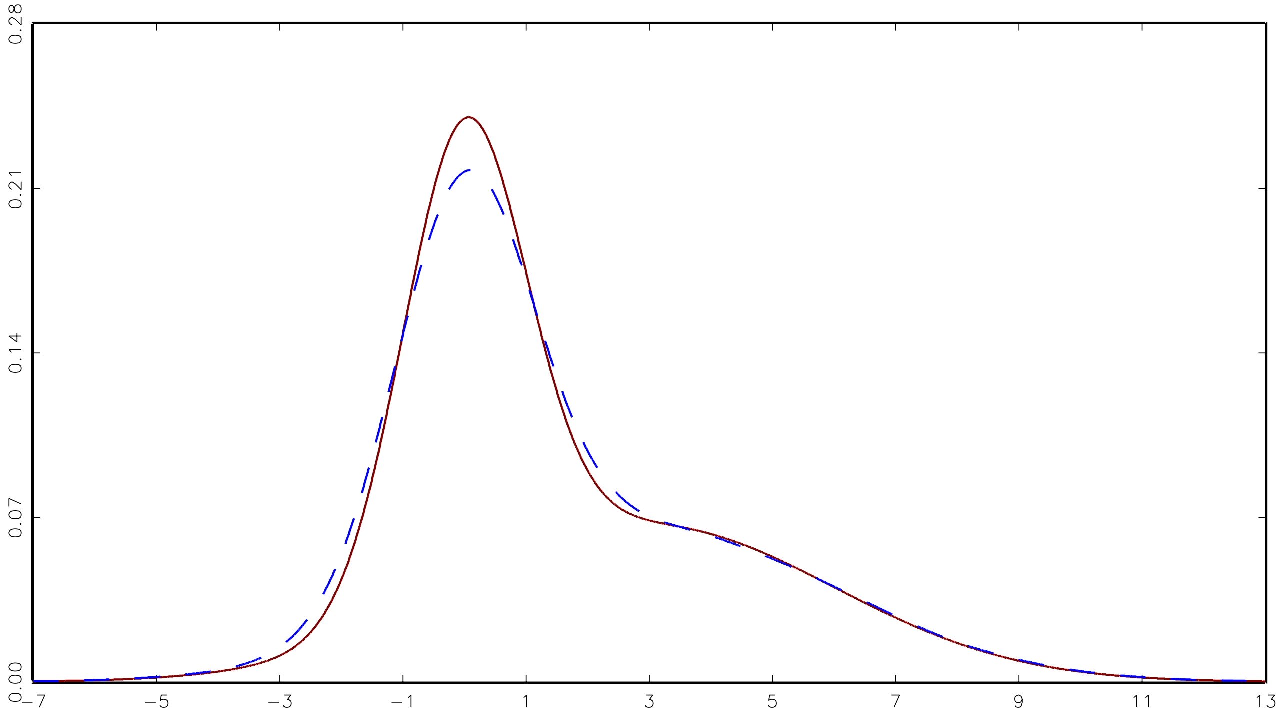


Figure12b

Unweighted estimate using h^*

Non−proportional Sampling n2/n1=2

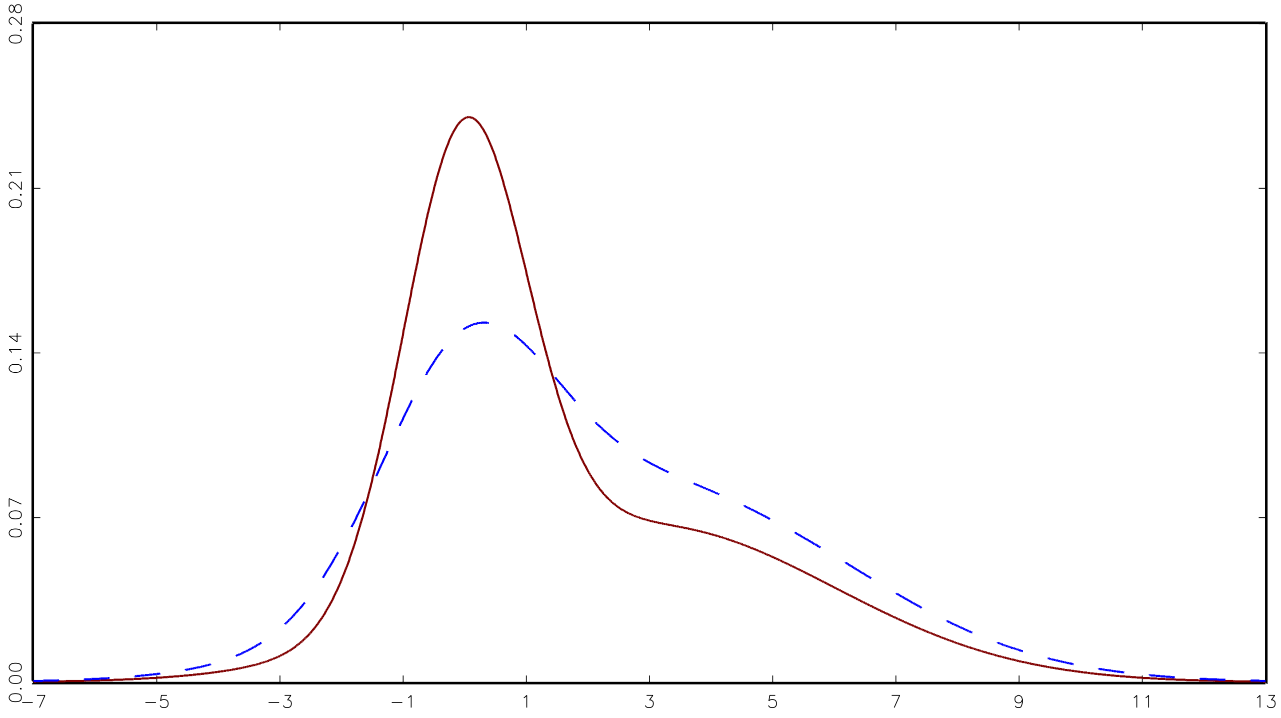


Figure12c
Unweighted estimate using h_a
Non−proportional Sampling   n2/n1=2

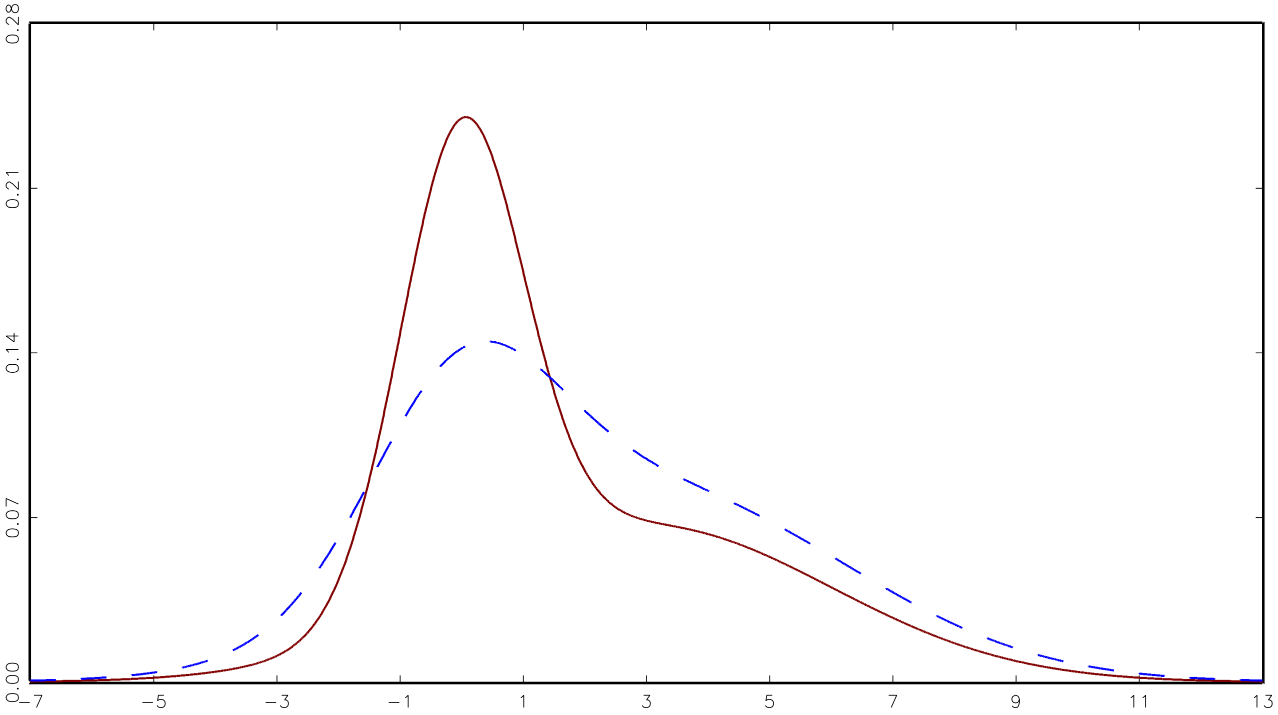
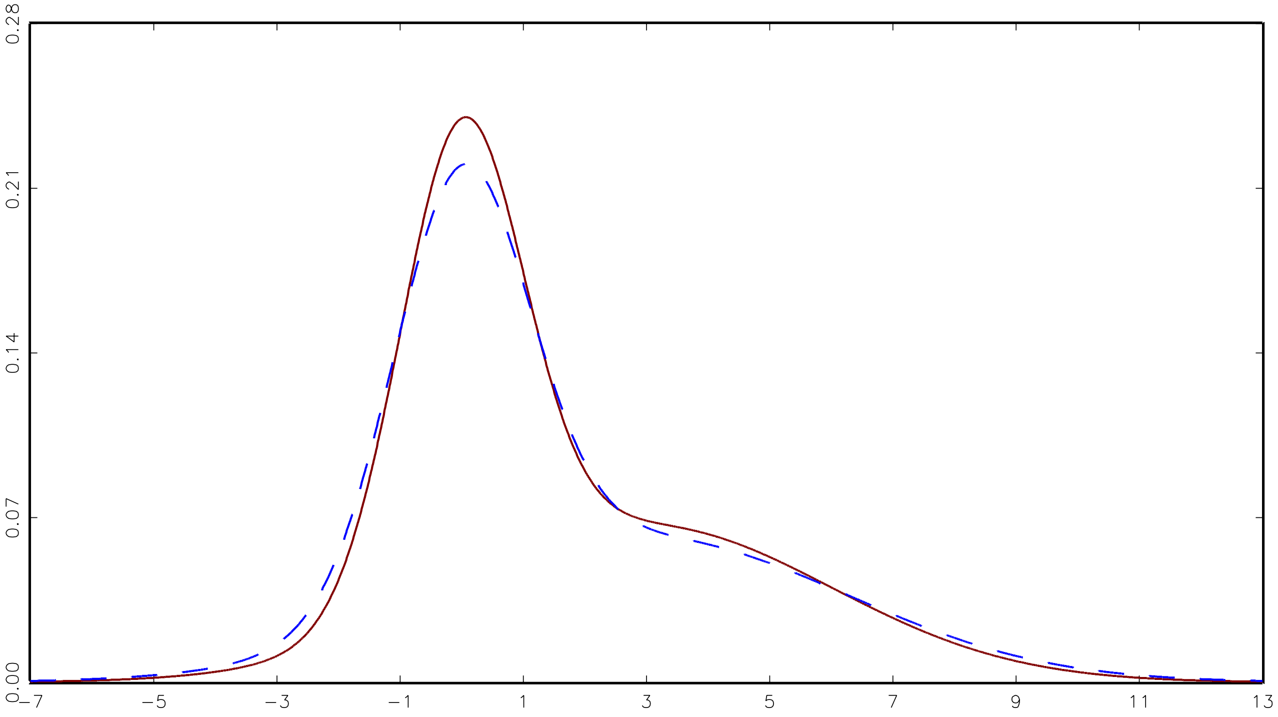


Figure12d
Weighted combination of stratum−specific densitiies
Non−proportional Sampling   n2/n1=2


0.28
0.21
0.14
0.07
0.00
−7
−5
−3
−1
1
3
5
7
9
11
13