Session Number: 4C
Session Title: Improving Estimates from Survey Data
Session Organizer(s): Stephen Jenkins, Holly Sutherland
Session Chair: Stephen Jenkins

Paper Prepared for the 29th General Conference of
The International Association for Research in Income and Wealth

Joensuu, Finland, August 20 – 26, 2006

The Importance of Cutting Corners: Estimating Robust Estimates for Grouped Income Data

Nicholas Biddle and Boyd Hamilton Hunter

For additional information please contact:

Author Name(s) : Nicholas Biddle
Author Address(es) : Centre for Aboriginal Economic Policy Research, Australian National University
Author E-Mail(s) : nicholas.biddle@anu.edu.au
Author FAX(es) :
Author Telephone(s) : 61-2-6125-8206

This paper is posted on the following websites: http://www.iariw.org

The importance of cutting corners: Estimating robust medians for grouped income data

Nicholas Biddle[1] and Boyd Hamilton Hunter[2]

Centre for Aboriginal Economic Policy Research

The Australian National University

For many applications, analysts are required to use data in ranges either because continuous data was not collected or because it was not made available to the researcher. Especially if the underlying variable is skewed, then calculations based on grouped data are likely to be influenced by the assumptions one makes regarding the distribution of values within the ranges. This paper summarises and tests a practical 'short-cut' for estimating medians using grouped data that takes into account more information than the standard method used by most statistical agencies and applied researchers.

Monte Carlo simulation methods are used to illustrate the relative bias in estimated medians using this 'short-cut' compared to using a proportional allocation of data within ranges. This 'short-cut' has lower mean squared error than the method used by most applied researchers. Real-world data are used to illustrate how trends in median income differ between using these two methodologies. Continuous survey data are also used to test the relative performance of the competing estimators.

[1] PhD Research Student

[2] Fellow

Introduction

In any analysis, the treatment of grouped income data is problematic in that it is not obvious what assumptions should be made to summarise the overall distribution. A common alternative to calculating a mean is to estimate the median, which is thought to be a more robust measure of central tendency. However, the estimates of median are themselves not independent of distributional assumptions as it is necessary to make suppositions about how to allocate respondents within the income range in which the median lies. This paper analyses the sensitivity of median estimates and proposes a practical method that can easily be applied by policy makers.

This paper outlines a practical 'short-cut' for estimating medians using grouped data, introduced in Altman, Biddle and Hunter (2004). We then use simulated data and Monte-Carlo simulation methods to illustrate the potential bias in estimated medians using the proportional allocation within income ranges to estimate the true median. The next two sections quantifies the difference between our short-cut technique and the conventionally estimate of median that fails to use any distributional information. Long run income trends for Indigenous Australians are first examined, followed by an analysis of the performance of the two estimators using continuous income data from a recent cross-sectional survey. While the difference is reasonably small it does affect the emphasis on the interpretation of the patterns in

medians. Consequently, the final section reflects on policy implications of using various estimates of medians.

A practical short-cut for estimating medians

The conventional method for calculating medians in many official publications is the proportional allocation of people within the relevant income ranges. That is, each dollar unit within each of the ranges is expected to contain the same number of people as the other dollar units within the same group. This assumed probability density function (pdf) is equal to the number of people within the group divided by the size of the interval.

While this uniform allocation method is likely to provide a reasonable estimate when the median is in a flat or symmetrical part of the income distribution, it will provide a biased estimate when the income distribution is highly skewed. Groups that include many welfare recipients are likely to be heavily skewed to the left and hence the median may reside in portion of the income distribution with a significant negative slope. This technique was first developed for Indigenous Australians, however it is equally applicable to other minority groups and pension recipients. The medians based on the proportional allocation of income ranges will tend to overestimate median incomes for such disadvantaged groups. It is particularly important to take into account the shape of the income distributions

when comparing very different groups such as Indigenous and non-Indigenous populations.

The remainder of this section outlines a method to estimate the shape of the distribution and hence the median income. It begins by defining some of the preliminary concepts, then outlines the simple four-step procedure introduced in a 2004 article by Altman, Biddle and Hunter (hereafter ABH) to estimate of the median.

The shape of the income distribution is represented by the probability density function (pdf), a standardised measure of the income distribution (i.e. representing the probability (Y=y)). The probability of having an income less than a particular income, y, is provided by the cumulative density function (cdf), which measures the area under the pdf curve up to y.

Normally, for the empirical researcher, income information is provided only in ranges so we only have information on the cdf at the boundaries of the income ranges (Y4, Y1 etc in Fig. 1). Furthermore, we do not know the pdf at any point along the distribution. Instead, we approximate the empirical pdf as a piecewise linear function that passes through the mid-points of the various income ranges. This is obviously a rough approximation of the true pdf, which in all likelihood would be a smooth function of income. The cdf can then be estimated using integral calculus as the area under the pdf. Although this method of interpolating medians was defined using calculus methods, the

remainder of this section provides an easy to follow geometric representation.

[place Figure 1 here]

Step 1: Shape of the pdf

As mentioned, the first step in estimating medians is to estimate the shape of the distribution. This paper assumes that the slope of the pdf within each group is determined by the empirical pdfs on either side of the category where the median is known to lie. To do this, we initially assume that the height of the pdf at the midpoint of each of the three categories (H1, H2 and H3) is the empirical probability of being in each respective group, divided by the number of units in that group (i.e. the estimated probability per dollar unit). That is:

$$
\begin{aligned}
H_1 &= \frac{P\left(Y_1 \le Y \le Y_3\right)}{\left(Y_3 - Y_1\right)} \\
H_2 &= \frac{P\left(Y_3 \le Y \le Y_5\right)}{\left(Y_5 - Y_3\right)} \\
H_3 &= \frac{P\left(Y_5 \le Y \le Y_7\right)}{\left(Y_5 - Y_7\right)}
\end{aligned}
\tag{1}
$$

Using these heights at the midpoints, the gradient of the pdf between Y_3 and Y_4 (g_1) and the gradient of the pdf between Y_4 and Y_5 (g_2) can be calculated as follows:

$$g_1 = \frac{H_2 - H_1}{Y_4 - Y_2}$$
$$g_2 = \frac{H_3 - H_2}{Y_6 - Y_4} \qquad (2)$$

For the Indigenous population, the median is likely to lie to the right of the mean, and hence these gradients are likely to be negative.

Step 2: Height of the pdf

Now that we have an estimate of the gradient for the pdf within the median group, the next step is to find the height of the pdf. Given we are assuming a linear pdf with a constant and known gradient, it is sufficient to know the height of the pdf at the lower and upper bounds of the category (and hence the height at the midpoint). To find the heights at these two bounds, we exploit the fact that we know the actual probability of being in that group, or the area under the curve. We label this known probability P. Given that the probability of being in that group is equal to the area under the pdf, we also know that the

probability is equal to the area to the left of the midpoint, plus the area to the right.

[place Figure 2 here]

Letting Δ equal the distance between the midpoint and both the upper and the lower bound of the category, $(Y_4 - Y_3)$ we now have the following equation:

$$P = \Delta(H_{mid}) + \frac{1}{2}\Delta(H_l - H_{mid}) + \Delta(H_u) + \frac{1}{2}\Delta(H_m - H_u) \tag{3}$$

Now because we know the gradient between H_l and H_{mid} is g_1, and the gradient between H_u and H_{mid} is g_2, we can also set up the following two equations:

$$\begin{aligned} H_l &= H_{mid} - g_1\Delta \\ H_u &= H_{mid} + g_2\Delta \end{aligned} \tag{4}$$

Putting Equation (4) into Equation (3) gives the following:

$$P = \Delta(H_{mid}) + \frac{1}{2}\Delta(H_{mid} - g_1\Delta - H_{mid}) + \Delta(H_{mid} + g_2\Delta) + \frac{1}{2}\Delta(H_m - H_{mid} - g_2\Delta)$$

Which can be solved to give:

$$P = \Delta H_{mid} - \frac{1}{2}\Delta^2 g_1 + \Delta H_{mid} + \Delta^2 g_2 - \frac{1}{2}\Delta^2 g_2$$

or

$$P = 2\Delta H_{mid} + \frac{1}{2}\Delta^2 \left(g_2 - g_1 \right)$$

That is:

$$H_{mid} = \frac{2P + \Delta^2 \left(g_1 - g_2 \right)}{4\Delta} \tag{5}$$

Using the height of the midpoint and the assumed gradients, we can now estimate the pdf at any point within that group. Furthermore, Equation (5) can be put back into Equation (4) to get values for H_l and H_u.

Step 3: Is the median to the left or the right of the midpoint?

Now that we have an estimate for the pdf, the next step in calculating the median is establishing whether it is to the left or the right of the midpoint of the income group. To do so, we estimate the cdf at the midpoint (which is the area under our estimate for the pdf) and see whether it is greater or less than 0.50.

The cdf of the midpoint, C_{mid}, is given by the following formula:

$$C_{mid} = C_l + \Delta H_{mid} + \frac{1}{2}\Delta\left(H_l - H_{mid}\right) \tag{6}$$

where C_l is the cdf up to but not including the income group that the median is in. So if $C_{mid} > 0.5$ then we know that the median is to the left of the midpoint, whereas if $C_{mid} < 0.5$, we know that it is to the right.

Step 4: Estimating the median

Now that we know at what part of the income group the median will be estimated to lie, we can now estimate where exactly the median is, based on our estimated pdf. As mentioned previously, this median is estimated differently if it is to the left of the midpoint as opposed to the right. This can be shown by the following diagram which shows the different ways in which the median is calculated

[place Figure 3 here]

Where we know the median is to the left of the midpoint, we know that the median is that value of income where the lightly shaded area in Figure 2 is equal to the difference between 0.5 and the cdf at the lower bound. Letting: Y_{med} equal the estimated median; δ the difference

between the estimated median and Y_3; and H_{med} the height of the pdf at the median, we know that:

$$C_l + \delta H_{med} + \frac{1}{2}\delta\left(H_l - H_{med}\right) = 0.5 \tag{7}$$

We also know that:

$$H_{med} = H_l + g_1\delta$$

As such:

$$0.5 - C_l = \delta\left(H_l + g_1\delta\right) + \frac{1}{2}\delta\left(H_l - H_l - g_1\delta\right)$$

or

$$g_1\delta^2 + 2H_l\delta - 2\left(0.5 - C_l\right) = 0$$

Solving this quadratic gives:

$$\delta = \frac{-H_l \pm \sqrt{H_l^2 + 2g_1\left(0.5 - C_l\right)}}{g_1} \tag{8}$$

Similarly, if the median is to the right of the midpoint, we have:

$$\delta = \frac{-H_{med} \pm \sqrt{H_{med}^2 + 2g_2\left(0.5 - C_{med}\right)}}{g_2} \tag{9}$$

The difference between Equations (8) and (9) is that to the right of the midpoint we use the height and estimated cdf at the midpoint $\left(H_{med}, C_{med}\right)$ rather than at the lower bound and we use the second gradient $\left(g_2\right)$ rather than the first.

Our estimated median is therefore either:

$$Y_{med} = Y_3 + \delta \tag{10}$$

or

$$Y_{med} = Y_{mid} + \delta \tag{11}$$

Whether we use Equation (10) or Equation (11) depends of course on whether we are to the right or the left of the midpoint.

Estimating the bias in medians using various techniques

Now that we have outlined the alternate method for estimating medians, it is important to see a) how close this estimator is to the true median and b) how well it performs relative to the proportional allocation method commonly used in empirical work. To do so, we first simulate some income data with a mean of $320, which was the mean weekly income for Indigenous Australians 15 and over in $2001 from the 2002 National Aboriginal and Torres Strait Islander Social Survey (NATSISS).[1] For our main results we simulated 100,000 observations, however we also compared the results to using 1,000 and 10,000 observations.

This income data was simulated using the gamma distribution in three ways to represent varying assumptions of skewness.[2] That is:

- Alpha = 1.2, Beta = 266.67;
- Alpha = 3, Beta = 106.67; and
- Alpha = 10, Beta = 32.

Estimates of the three pdfs for these distributions are given in Figure 4 below.

[place Figure 4 here]

Now that we have three sets of income distributions, the next step is to set up income groupings, keeping in mind that in practice, it is only the number of people in each of these income groups which the majority of applied researchers know. We use two income category breakdowns, one with 14 income categories that matches the income groupings in

the 2001 census and one with only seven groups. The lower bounds of these are:

- 14 categories: 0, 40, 80, 120, 160, 200, 300, 400, 500, 600, 700, 800, 1000, 1500; and
- 7 categories: 0, 100, 300, 500, 750, 1000, 1500.

For each combination of gamma distribution and income groupings, medians were then estimated using both the ABH technique and commonly used linear interpolation method. These estimated medians were then compared to the true median from the continuous distribution.

To compare the two estimators, we ran Monte Carlo simulations with 100 repetitions and within each repetition, bootstrapped the standard errors using 200 repetitions. Using the bias of the estimated median as well as the standard error, we generated the Mean-Squared Error (MSE) as:

$$MSE\left(\hat{Y}_{med}\right) = Var\left(\hat{Y}_{med}\right) + \left(Bias\left(\hat{Y}_{med}\right)\right)^2 \tag{12}$$

For more information on the calculation of biases and MSEs, see Greene (2000). The results from this exercise are given below in Table 1.

[place table 1 here]

Table 1 shows that on the one hand, when using the income groupings from the 2001 Census, the bias and MSE is smaller for the ABH

technique as opposed to the proportional allocation method. On the other hand, for the two most skewed distributions (alpha = 1.2 and alpha = 3), the standard error is higher for the ABH technique. Clearly, incorporating distributional information in an estimate usually has a small cost in terms of reducing the reliability of estimated medians. However, this cost tends to be outweighed by the benefit of having a substantially lower bias and MSE when using the ABH technique.

Using the broader income categories, the ABH technique has a lower bias and MSE for alpha = 1.2 and alpha = 10, but has a slightly higher values for alpha = 3. However, the differences between the ABH and proportional allocation techniques is relatively minor for alpha = 3 with the bias and standard errors of estimates being in the lower range for both estimators in Table 1. That is, increasing the breadth of income categories may reduce the efficacy of the ABH technique vis-à-vis proportional allocation, possibly because broader income categories render the distributional information less meaningful (e.g. when it spans diverse parts of the distribution—i.e. covering both increasing and decreasing portions of a pdf).

Long run trend in Indigenous income

The previous table has shown that, on balance, the ABH technique is a 'better' estimator of the true median using simulated data. An obvious question is how much difference does the use of ABH technique make in practice. To demonstrate that conclusions made from real data can

differ substantially when using the ABH and the conventional estimators, we present median income estimates from the Australian Censuses conducted in 1981, 1991 and 2001.

[place Table 2 here]

For the most part, the results are similar for both median estimators. For example, there was a 13 per cent increase in median income for Indigenous individuals over the 20 years between 1981 and 2001 when the ABH technique is used. However, there was a 20 per cent increase in median income for the same individuals when the proportional allocation technique was used. Looking at the ratio of Indigenous to non-Indigenous medians, the conclusion about the changes in relative disadvantage of Indigenous Australians depends on what method is used. While the ratio of medians was the same for both estimators in 1981, the ABH technique estimates that there was substantially less improvement in the relative income status between 1981 and 2001 than the more commonly used technique.

There is relatively little difference in the trends in median estimators for household income. Indeed, there is virtually no difference in the ratio of Indigenous to non-Indigenous medians for the respective estimators. One explanation for this is that raw household income for Indigenous households tends to be closer to (and more symmetric with) the non-Indigenous distributions because of larger size of many Indigenous households (Hunter, Kennedy & Smith 2003). However, the use of

equivalence scales on income data would increase the differences between the Indigenous and non-Indigenous distributions.[3] Notwithstanding, the contrast between the analysis of individual and raw household income illustrates that the ABH technique will only substantively change the results if the differences in the shape of the respective distributions are large.

Testing using the NATSISS

In 2002, the ABS undertook the National Aboriginal and Torres Strait Islander Social Survey (the NATSISS), which had continuous income data on 9,127 Indigenous Australians aged 15 years and over (from a sample of 9,359). While the sample sizes are not sufficient to estimate the income characteristics below the state/territory level, and there are no historical datasets to compare the results against, the NATSISS does allow us to test the ABH technique of estimating the median income of Indigenous Australians against the standard techniques.

Using continuous individual income from the NATSISS, median income is $230 (mean $335). To test the ABH method of estimating median income, we assume that instead of collecting continuous income, data was collected in the same 14 ranges used in the 2001 Census presented earlier, and the only output that was available was the proportion of people with an income in that range. We then re-estimate median income using both the ABH and proportional allocations methods, assuming that this was all the information available.

Both methods overstate median income. However, the difference between the calculation from the ABH method of $257.56 and actual median income, is less than the difference between the calculation from the proportional allocation method of $268.01. In summary, therefore, although using grouped income inevitably leads to a loss of information, by taking into account information on either side of the median income group, it is possible to get a closer estimate of median.

Concluding remarks

The main result from this paper has been to show that, when estimating median income from grouped data, assuming a uniform distribution within the ranges does not necessarily result in the closest estimate to the true value. Rather, we have outlined the ABH technique that is reasonably easy to implement (code is available from the authors on request) and, at least in the distributions we tested, almost always has a lower bias and a lower mean square error (only one exception in our simulations). Furthermore, using continuous income data from the 2002 NATSISS, the estimate using the ABH technique is closer to actual median income.

While the 'real' data example showed that the differences between estimators are usually small, the policy conclusion may vary substantially depending on which technique used. Given that the ABH technique uses the distributional information from the empirical data, and hence tends to cut the corners off the pdf based on the linear

interpolation of income ranges, it can be argued that it is important to 'cut corners' when estimating medians.

Our empirical example focused on the Indigenous population as they are one group in Australia for whom grouped data is generally all that is available, and whose median is likely to lie on a part of the distribution with a large (negative) slope. Other groups are also likely to have similar income distributions, for example, old-age pensioners or single mothers. Consequently, it would be appropriate to use the ABH technique when examining income data for these groups.

Although continuous income is rarely available in the publicly available data collections, statistical agencies often have this information available to them. It would be useful for those with actual distributions to test how close the ABH technique comes to the true median, and compare this with the proportional allocation technique.

In the meantime, the evidence presented in this paper point to the clear superiority of techniques that use all the available information on the underlying income distribution. Hence where grouped data are the only viable source of data, the ABH or similar techniques should be used to estimate medians—especially where there are reasons to expect that the underlying income distributions are heavily skewed.

References

Altman, J.C., Biddle, N. and Hunter, B.H. 2004. 'Indigenous socioeconomic change 1971-2001: A historical perspective', CAEPR Discussion Paper No. 266, CAEPR, ANU, Canberra.

Armitage, P., Berry, G. and Mathews, J.N.S. 2002. Statistical Methods in Medical Research, Blackwell Science, Massachusetts.

Evans, M., Hastings, N. and Peacock, B. 1993. Statistical Distributions, John Wiley & Sons, New York.

Greene, W.H. 2000. Econometric Analysis, Prentice Hall, New Jersey.

Hunter, B.H., Kennedy, S. and Biddle, N. 2004. 'Indigenous and other Australian poverty: Revisiting the importance of equivalence scales', Economic Record, 80 (251): 411-22.

Hunter, B.H., Kennedy, S. and Smith, D. 2003. 'Household composition, equivalence scales and the reliability of income distributions: Some evidence for Indigenous and other Australians', Economic Record, 79 (244): 70-83.

Table 1. Bias, variance, and MSE of median estimators using gamma distribution

		Estimated median	Bias of estimator	Standard error of estimator	MSE of estimator
14 income categories					
Alpha = 1.2	ABH technique	12,324	14	142	20,391
	Proportional allocation*	12,502	164	138	46,045
Alpha = 3	ABH technique	14,843	7	95	9,025
	Proportional allocation	14,896	46	87	9,626
Alpha = 10	ABH technique *	16,141	43	66	6,266
	Proportional allocation*	16,191	92	68	13,186
7 income categories					
Alpha = 1.2	ABH technique *	12,479	142	125	35,756
	Proportional allocation*	12,925	587	106	355,952
Alpha = 3	ABH technique	14,864	14	90	8,219
	Proportional allocation	14,859	9	89	7,969
Alpha = 10	ABH technique *	16,355	257	91	74,270
	Proportional allocation*	16,418	320	94	110,969

Notes: Monte Carlo simulations reported in this table were conducted for 100,000 observations. Simulations were also conducted for 1,000 and 10,000 observations with the pattern of results almost identical with those in this table. The obvious exception is that the variance of the estimators is higher with smaller samples. The mean was held constant at $16,640 which was the average annual income for Indigenous Australians in the 2002 National Aboriginal and Torres Strait Islander Social Survey). An asterisk denotes that the estimator was significantly different from the true median at the 5 per cent level.

Table 2. Annual median individual and household income (in $2001), 1981–2001

Variable	1981	1991	2001
Individual income: Indigenous			
ABH technique	9,750	10,972	11,055
Proportional allocation	9,818	11,284	11,760
Individual income: Non-Indigenous			
ABH technique	17,732	17,784	19,744
Proportional allocation	17,771	17,877	19,818
Ratio of Indigenous to non-Indigenous medians			
ABH technique	0.55	0.62	0.56
Proportional allocation	0.55	0.63	0.59
Household income: Indigenous			
ABH technique	35,178	33,961	40,929
Proportional allocation	35,416	34,117	40,954
Household income: Non-Indigenous			
ABH technique	48,760	44,387	52,510
Proportional allocation	48,709	44,386	52,598
Ratio of Indigenous to non-Indigenous medians			
ABH technique	0.72	0.77	0.78
Proportional allocation	0.73	0.77	0.78

Figure 1. Defining preliminaries for median calculations

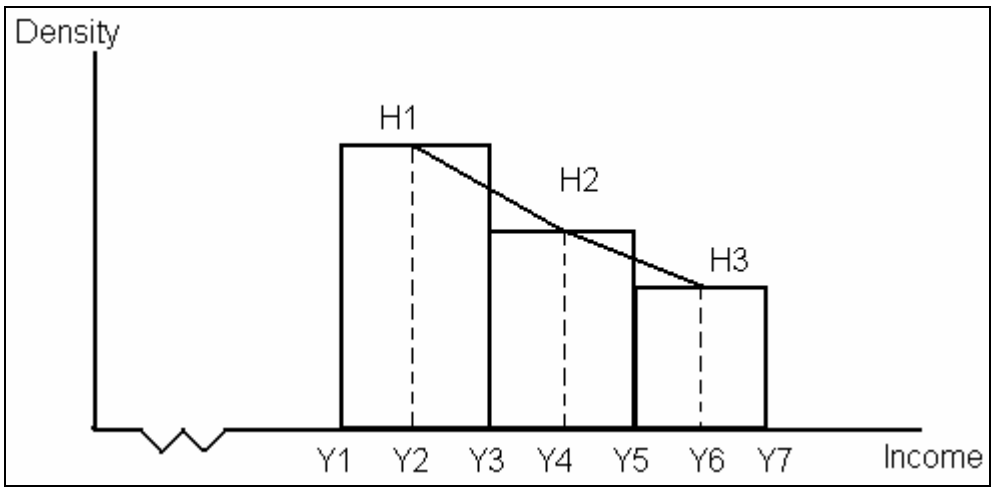


Figure 2. Height of pdf

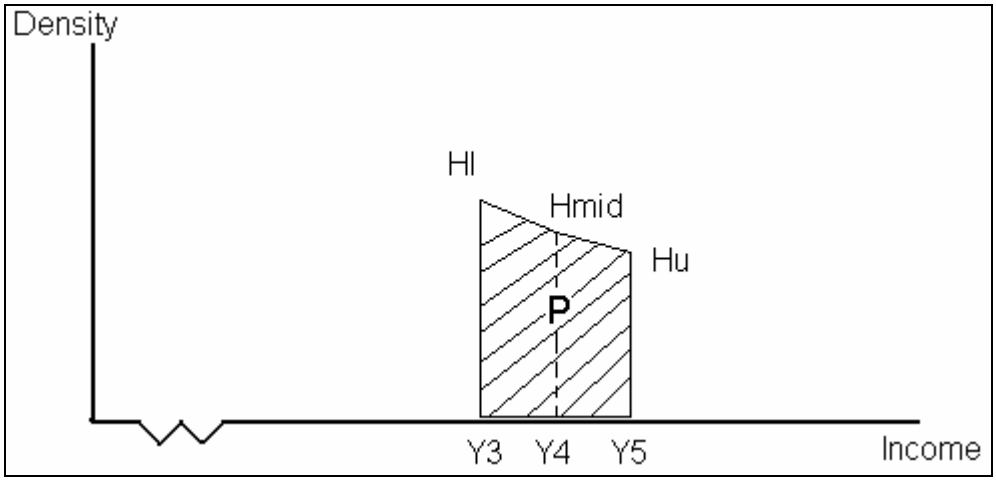


Figure 3. Location of median

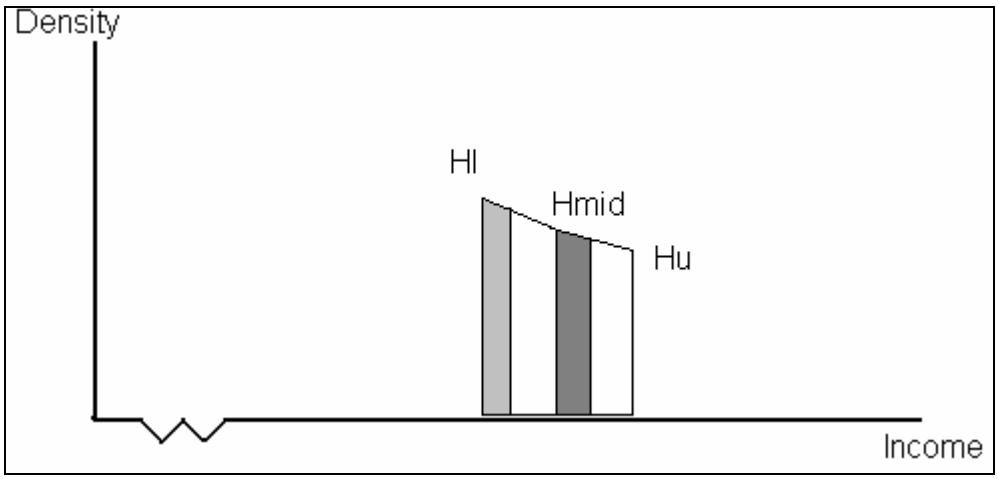
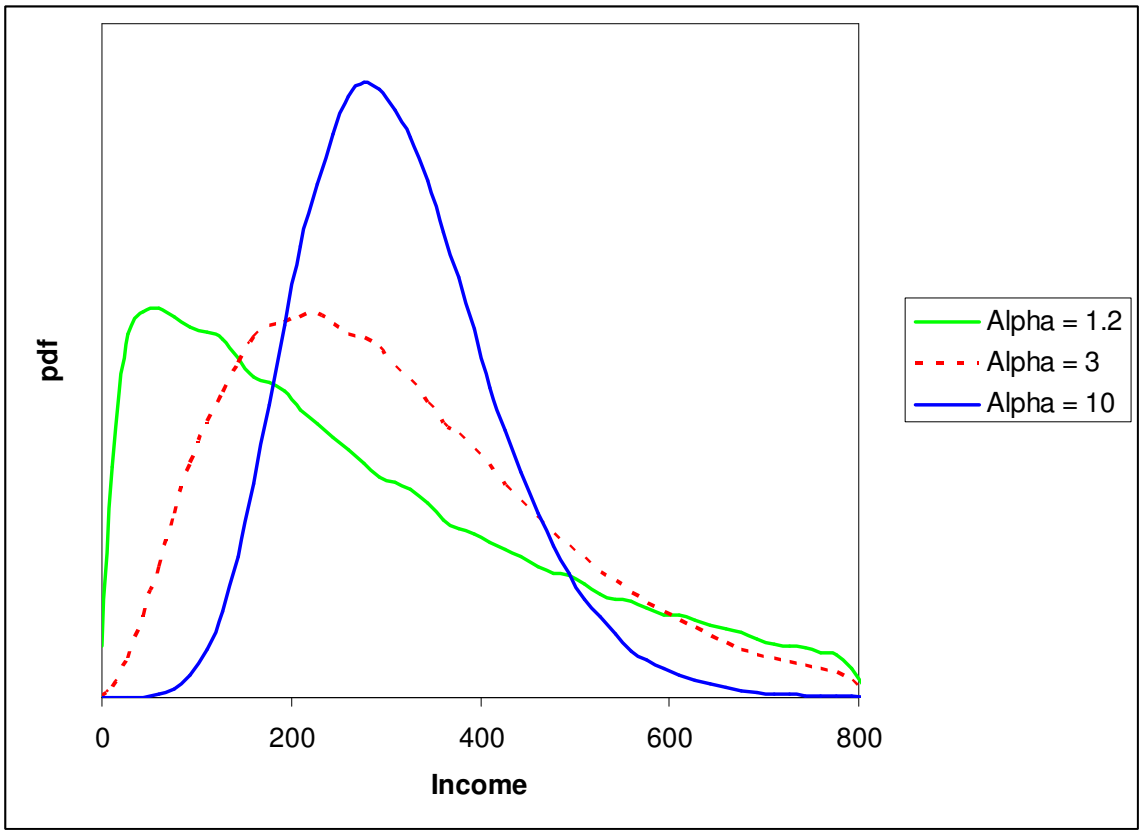


Figure 4. Simulated distributions



Note: The probability density function for these graphs were generated using the 'kdensity' procedure in Stata Version 8 with the seed set to 48901

Notes

1. We used $2001 because we were testing the results against the income categories used in the 2001 Census of Population and Housing.

2. For more information on the gamma distribution, please see Armitage, Berry and Mathews (2002). The parameterisation of the gamma distribution and the relationships between the various parameters and the moments of the distributions are clearly layed out in (Evans, Hastings & Peacock 1993: 75-81)

3. It is rather difficult to estimate equivalent income using grouped income data from the Census. In any case, the use of equivalence scales adds an extra dimension of error into the estimates that would confound the interpretation (Hunter, Kennedy & Biddle 2004).