Comparison of personal income inequality estimates based on data from the IRS and Census Bureau

Kitov, Ivan

IDG RAS

18 October 2007

Comparison of personal income inequality estimates based on data from the IRS and Census Bureau

Ivan O. Kitov

Introduction

In professional economic literature, paper and electronic media, and numerous blogs devoted to economics one can observe an intensive ongoing discussion of increasing economic inequality in the USA. For example, a prominent economist Paul Krugman explicitly articulated that economic inequality in the USA is one of the favorite topics in his blog – "The Conscience of a Liberal". His blog belongs to a family representing various aspects of economic inequality as measured by personal and household income distribution.

There is a strong difference between Krugman's (and other economists') and my quantitative assessment of personal income inequality, however. Using personal income distributions, which have been reported since 1947 by the US Census Bureau (2007), I conducted quantitative estimates of Gini coefficient and found that this coefficient was practically constant over time (Kitov, 2007). Having a constant Gini coefficient since (at least) 1947, one might find it strange that other researches and media thoroughly discuss increasing inequality during the last 25 years. It was difficult to actually understand why those researches do not use the US Census data despite the Census Bureau (2004) explicitly states:

> Because of its detailed questionnaire and its experienced interviewing staff trained to explain concepts and answer questions, the CPS ASEC is the source of timely *official* national estimates of poverty levels and rates and of widely used estimates of household income and individual earnings, as well as the distribution of that income.

Krugman explained why he and other researchers are forced to deny the estimates based on Census Bureau data. In his relatively old post, On Tracking Inequality, he gave some details of his approach to income distribution:

> First, because Census data are based on a limited sample, not the whole population, they're unreliable in tracking the income of small groups – and the really rich are a small group, who just happen to bulk large in the economy. Second, the questionnaire is "top-coded": if the individual interviewed has earnings higher than $999,999, those earnings are recorded simply as $999,999.

> Since a lot of income growth in the last few decades has taken place among people with multimillion-dollar incomes, the Census data miss an important part of the story.

In practical and theoretical terms, both statements (reasons) are wrong. First, in hard sciences, one is not often able to measure true values of desired variables, but usually measures some portions of them. For example, nobody tries to invent a weighting machine in order to measure the Earth's mass. It is enough to measure gravity acceleration in one point since this acceleration is proportional to the total mass. Therefore, if a portion of a whole object is representative one and is measured consistently over time, one can carry out a reliable quantitative analysis. A randomly changing portion, as sometimes happens to macroeconomic variables after introduction of new definitions, would, obviously, ruin any such quantitative analysis. So, using surveys of small population samples does create a problem with internal precision, but should not necessary disturb results of overall quantitative analysis. Kitov (2007) provided quantitative results which confirm that the Census Bureau has been collecting high-quality data.

Second, the "top-coded" approach does not harm the estimates of income in the "the richest of the rich" group. This effect is known more than hundred years already. Higher incomes are very accurately distributed according to the Pareto (power) law. As shown below, the tax-based (gross) income distributions in the USA confirm this observation. As a matter of fact, one does not need to measure any personal income in the high-income group. S/he needs to estimate the number of persons with income above some given (high) threshold. Then, one can use simple mathematical equations to obtain accurate population density at any income level and also total income above any threshold.

What is then the problem with the IRS based measures of income, which result in changing inequality? This paper is aimed at the development of a simple answer - these inequality measures are based on income definitions allowing floating low-end income threshold. In other words, population used for inequality calculation fluctuates randomly or according to some predetermined relationship.

The effect of changing population basis due to numerous revisions of income definition is also observed in Census income data (Census Bureau, 2003). The portion of

people with income severely changes over time. It was increasing in the 1960s and 1970s due to a strong growth in women's participation rate. It has been falling since 1990, however. When people without income are included in calculations of income inequality, the Gini coefficient (for personal incomes) actually has been intensively falling since 1947 due to a strong growth of the portion of people with income. So, one can conclude that the driving force behind the increasing personal income inequality, as reported by the IRS, likely consist in biased measurements and inconsistent definitions.

It is of principle importance for the current study that despite the changing income definition and corresponding population basis the estimates of income inequality were not changing in the group with non-zero income. This observation contradicts the changing inequality as obtained from the IRS data. Only quantitative analysis can resolve this conflict. The resolution of the conflict is the purpose of this paper. Because the results showing increasing inequality are quantitative, it is feasible to exactly show the reasons for the observed contradiction and indicate caveats in Krugman's (and other's) approach.

IRS and Census Bureau inequality estimates

Original (real gross) income distributions are reported by the IRS (2007). Table 1 provides the numbers of people in predefined income bins (in chained 1990 dollars) for 1990 and 2004. Also listed are widths of income bins, which are used below for calculations of population densities, and centers of income intervals. The income bins are fixed over time and not adjusted for the growth of real economy and in working age population. The lowermost income bin contains zero income and net loss reports. The highest income bin includes those income reports which exceed $10,000,000. This is an open-end income bin without the estimate of average income. Fortunately, only several thousand people have incomes above $10,000,000. First thousands is not the number which could influence much the overall income inequality estimate. Moreover, these richest people also distributed according to the Pareto law, i.e. measured and estimated total income in this bin should not differ much.

The IRS income tables provide a basis for current estimates of economic inequality in the USA. Conventional conclusion about income inequality is very

consistent among economists – the inequality has been rising during the last 20 years. At first glance, this conclusion is quantitatively correct, but I will argue that it is wrong due to potential inaccuracy in methodology and unacceptable misinterpretation of quantitative results.

Figure 1 compares (gross) income distributions for 1990 and 2004, as listed in Table 1. Since the income bins presented in Table 1 are of increasing width one can observed some spikes in the distributions. These spikes are, obviously, related to those income bins, which are wider than their predecessor. For example, the bin between $25,000 and $30,000 ($5000-wide) is followed by the bin between $30,000 and $40,000, ($10000-wide). Therefore, one can expect a larger number of people in the latter bin than in the former one. This effect is clearly observed in Figure 1, where the enumerated populations are assigned to the centers of corresponding income bins. Here and below, I prefer to use log-log scale in order to present highly changing population (and density) distributions in a very wide range of income, spanning seven orders of magnitude. The lowest income bin, corresponding to zero and negative (loss) reported incomes, is artificially associated with $1 income. The bin with incomes above $10,000,000 is not shown because of the absence of mean income estimate in this bin.

One can easily derive an obvious conclusion from Figure 1- there are more people with lower, middle and high incomes in 2004 than in 1990. This is a mechanical result of increasing population – more and more people get income as working age population is growing.

One should normalize the curves to total population (with reported income) in given years in order to obtain population independent results. In addition to this normalization one can use population density instead of original population estimates in width changing bins. Income bin width would not be a problem for constant widths. Therefore, when the measured populations are normalized to corresponding income bin widths one obtains density of population as a function of income, i.e. the number of people per $1 bin. As before, we assign the obtained population densities to the centers of corresponding bins. The assignment of the density readings to the centers can potentially to disturb the observed curves when income bins are very wide and income distribution is described by a power law (Kitov, 2007).

Figure 2 depicts the population density curves obtained after the normalization of the curves in Figure 1 to the total population with (IRS reported) income, which includes people without income and those with incomes above $10,000,000, and to widths of corresponding income bins. As expected, both curves accurately follow the Pareto law distributions, which are represented by straight lines in log-log coordinates. This allow simple theoretical consideration of the distributions, as mentioned in Introduction. The most prominent features of the obtained curves are the increasing deviation between them staring from $62,500 (1990 dollars) and the fact that they are practically indistinguishable below this income threshold. As a rule, modern studies of income inequality find their conclusions in these population density curves. The curves, apparently, evidence that the portion of population with higher incomes has been growing since 1990 and as a consequence the economic inequality has been increasing.

This is not the end of the story, however. There is one question left. What is the effect of increasing total personal income on the observed population density distribution? Actually, total personal income grew from $3.41E+12 to $4.70E+12 (1990 dollars) between 1990 and 2004. So, the larger total income is a possible reason for the increased number of people with higher incomes. Then the same level of population density at lower incomes might be an artifact associated with inaccurate measurements at very low incomes or exclusion of some categories of income from IRS definition. This can be a big problem for the compatibility of estimates over time, as the Census Bureau discusses in methodological documents (US CB, 2003).

What does really happen when dimensionless (or relative) income distribution is used instead of that obtained in absolute income values? Two curves in Figure 3 represent those in Figure 1, which are additionally normalized to total personal income reported by the IRS, i.e. to $4.70E+12 in 2004 and $3.41E+12 in 1990. Income scales in 1990 and 2004 are also normalized to these total incomes and represent now dimensionless portions of total income. (For example, $10,000 in 1990 is transformed into (1.0E+4/3.41E+12 =) 2.93E-9 in 1990 and into 2.13E-9 in 2004.) As a result, widths of the given income bins in 1990 and 2004 also become different since relevant income scales are compressed by different factors. Also, the centers of original income bins

which were the same in 1990 and 2004 are now shifted relative to each other. In Figure 3, the curves in Figure 2 are compressed by different factors and shifted against each other.

The curves now represent population density as a function of dimensionless income and practically coincide at higher incomes and diverge at low incomes. Therefore, the density of population at higher incomes, as measured in dimensionless portions of total income, is practically the same in 1990 and 2004, considering the effectiveness of the IRS work and possible measurement errors. In other words, rich people have the same (within the uncertainty of income measurements) portion of the total income pie. In relative terms, these high-income people in 2004 are not richer than in 1990.

In the low income zone, the distributions are diverging with time. There are several explanations of this observation. First, this is the results of some real (objective) processes of income redistribution between rich, middle class and poor people in the USA. This is a common opinion in economic literature and media. Because of the changes in the measured personal income distributions one needs some driving force explaining the process. Second reason for the changing distribution is not related to increasing income inequality but is associated with lower (and varying) accuracy of income measurements at smaller incomes (possibly driven by definitions).

In the case of actual income redistribution process, one can expect some consistency between measures of income inequality provided by different agencies. For example, inequality estimates provided by the US Census Bureau, which include many taxable income sources and some extra sources as well, would be expected to confirm the IRS results. This is not the case, however, as mentioned before.

Figure 4 presents population density distributions obtained from Census Bureau income data according to the procedure described in Section 1 for the same years. There is no significant difference between distributions in 1990 and 2004. This result is confirmed by a wider set of observations between 1947 and 2005. There is no significant change in empirically determined Gini coefficient after 1960, as Figure 5 demonstrates. A slight decrease in Gini before 1960 is likely driven by data resolution problem (Kitov, 2007). Therefore, the reason for the discrepancy reported by the IRS might be associated with income definitions and reports.

So, there is a conflict between quantitative estimates based on the IRS and Census Bureau data sets. Which measure is a more reliable one? Let's consider two aspects of relevant income distributions – population basis and total personal income reported by the IRS and Census Bureau. It is likely that larger portions of working age population and real GDP potentially provide more reliable estimates of inequality.

Figure 6 presents the evolution of the portion of working age population with income as reported by the IRS and Census Bureau between 1990 and 2004. The number of people with IRS reported income is about 113,000,000 in 1990 and 132,000,000 in 2004. The Census Bureau reported ~181,000,000 in 1990 and 205,000,000 in 2004 from total working age population of ~194,000,000 in 1990 and ~230,000,000 in 2004. Corresponding portions are 0.93 and 0.58 in 1990, and 0.89 and 0.57 in 2004 reported by the CB and IRS, respectively. Therefore, the CB surveys cover a larger portion of population with income measurements.

Moreover, the surveys include taxable incomes as a subset of all measured incomes. Figure 7 illustrates the differences in income definitions between the IRS and CB. Gross personal income measured by the CB is of 70 per cent of real GDP – falling from 73% in 1990 to 67% in 2004. At the same time, the IRS reports only from 58% of real GDP in 1990 to 54% in 2004. It is also important that the IRS curve is a much higher volatility. This observation is potentially related to changes in (taxable) income definition.

Apparently, the IRS covers smaller portion of population and gross personal income than the Census Bureau. Basically, the IRS reports some income subset relative to the Census Bureau. Therefore, the observed difference between economic inequality estimates based on IRS and Census Bureau data is likely results from lower reliability of the IRS estimate. Income reports can not provide a consistent measure of personal income.

Conclusion

This paper quantitatively demonstrates that modern estimates of income inequality based on the data reported by the IRS are not reliable. The principal problem of the estimates is

highly volatile incomes of people in the low-end of income distribution. This volatility is likely related to measurement errors, changes in definitions or improper reporting.

IRS income estimates at high and the highest incomes are robust and follow the Pareto law. When normalized to total population with income and total (gross) personal income personal income distributions for 1990 and 2004 practically coincide. Hence, the inequality estimates based on the IRS data are distorted by reading in the low-income zone.

Income data provided by the US Census Bureau are consistent over time in all income ranges. Results presented by Kitov (2007) demonstrate that personal income distributions based on readings obtained in the Current Population Survey are characterized by practically constant Gini coefficient since 1960. This observation implies that normalized personal income distributions are also not changing with time.

References

Internal Revenue Service, (2007). Selected Income and Tax Items from Inflation-Indexed Individual Tax Returns. All Returns: Sources of Income, Adjustments, and Tax Items in Constant 1990 Dollars, http://www.irs.gov/taxstats/indtaxstats

Kitov, I., (2007). Modeling the evolution of Gini coefficient for personal incomes in the USA between 1947 and 2005, Working Papers 67, ECINEQ, Society for the Study of Economic Inequality, www.ecineq.org/milano/WP/ECINEQ2007-67.pdf

US Census Bureau, (2003). Technical Paper 63 Revised, Design and Methodology, http://www.census.gov/prod/2002pubs/tp63rv.pdf

US Census Bureau, (2004). Guidance on Differences in Income and Poverty Estimates from Different Sources, August 19, 2004

US Census Bureau, (2007). Current Population Reports Consumer Income Reports from 1946-2006 (P60), http://www.census.gov/prod/www/abs/income.html

Tables

Table 1. Personal income distribution according to the IRS

Income bin	Width	Center	1990	2004
No adjusted gross income		[1]	904876	1854886
$1 under $5,000	5000	2500	16478272	17039057
$5,000 under $10,000	5000	7500	14952855	17211889
$10,000 under $15,000	5000	12500	13922750	15889660
$15,000 under $20,000	5000	17500	11543228	13056490
$20,000 under $25,000	5000	22500	9572317	10990767
$25,000 under $30,000	5000	27500	7838225	8567162
$30,000 under $40,000	10000	32500	12282786	13309262
$40,000 under $50,000	10000	35000	8837067	9928723
$50,000 under $75,000	25000	62500	10944102	13635393
$75,000 under $100,000	25000	87500	3276142	4934480
$100,000 under $200,000	100000	150000	2329562	4213077
$200,000 under $500,000	300000	350000	644027	1211221
$500,000 under $1,000,000	500000	750000	130252	240876
$1,000,000 under $1,500,000	500000	1250000	29060	61800
$1,500,000 under $2,000,000	500000	1750000	11581	26977
$2,000,000 under $5,000,000	3000000	3500000	15331	39047
$5,000,000 under $10,000,000	5000000	7500000	3184	9625
$10,000,000 or more	>10000000		1522	5651

Figures

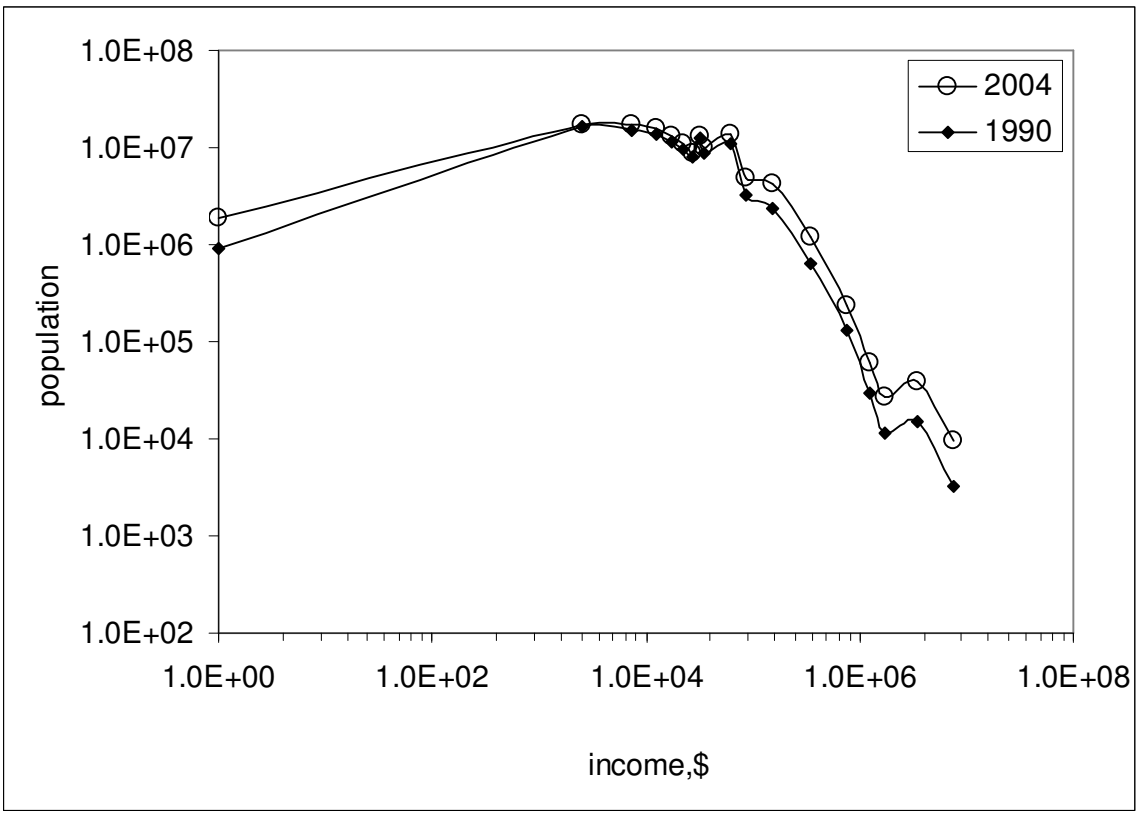

Figure 1. Comparison taxable income distribution, as reported by the IRS, in 1990 and 2004. Income bins are of increasing width. Enumerated populations are assigned to the centers of corresponding bins. Notice log-log scale. The lowest income bin corresponds to zero and negative (loss) reported incomes, i.e. people without income. The bin with incomes above $10,000,000 is not shown because of the absence of mean income estimate in this bin, i.e. x-value is not available.

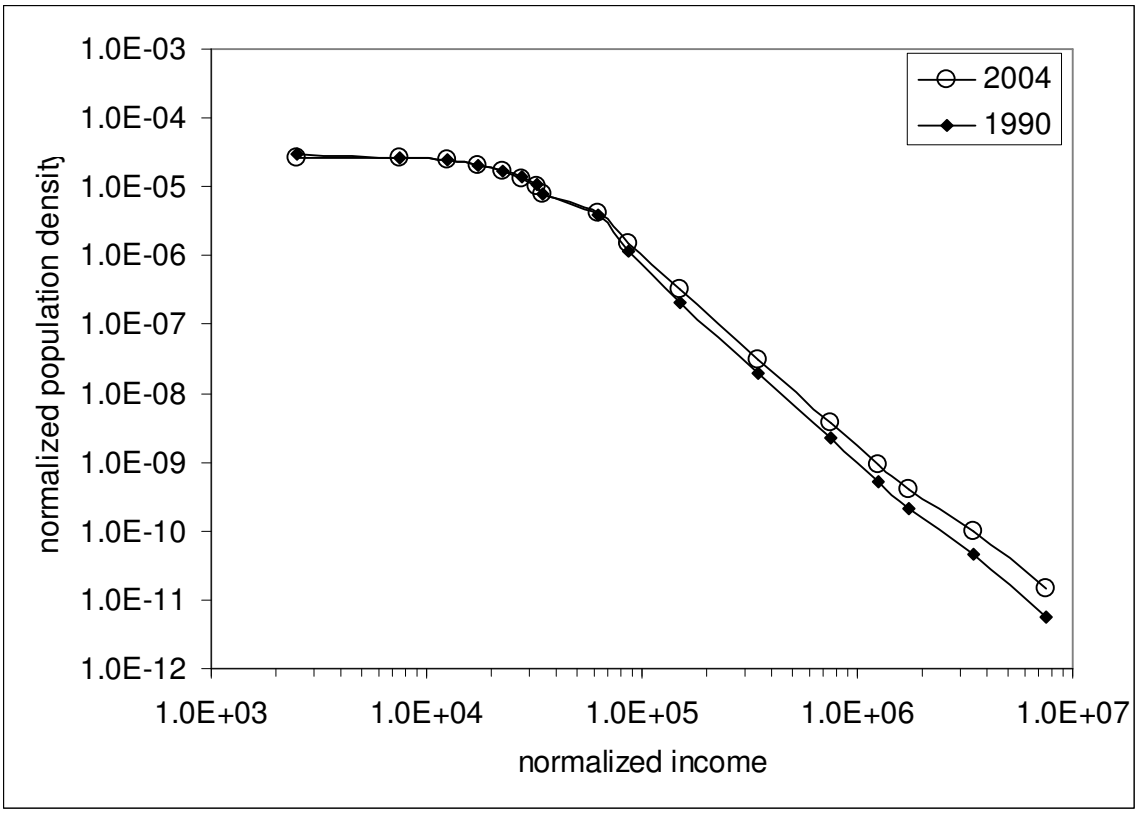

Figure 2. The curves in Figure 1 are normalized to total population with income reported to the IRS and to widths of corresponding income bins. Resulting population density distributions, i.e. the number of people per $1-wide bin, are plotted as a function of income (central point of corresponding income bin). First (zero width) and the last (open-ended) income bins are not presented. The curves almost coincide below $62.500 and then diverge with increasing income. As a result, income inequality seems to increase as the number of people with higher incomes increases faster than that with low incomes.

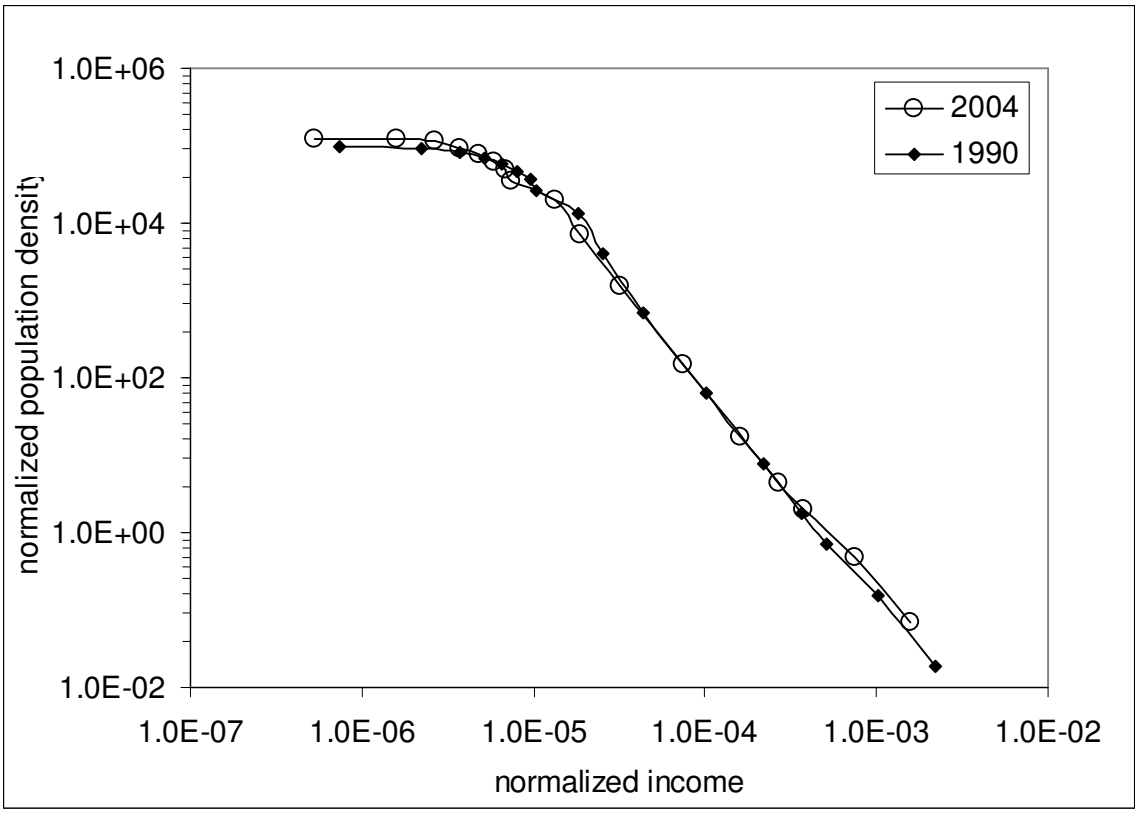

Figure 3. The curves in Figure 2 are additionally normalized to total personal incomes, i.e. to $4.70E+09 in 2004 and $3.41E+09 in 1990. The income scales are also normalized to these total incomes and represent dimensionless portions of total income. As a result, widths of the income bins also become different since the incomes scale in 2004 and 1990 are contracted by different factors. In turn, the centers of the same original income bins in 1990 and 2004 are shifted against each other. Effectively, the curves in Figure 2 are contracted by different factors and shifted against each other.

The curves now represent the density of population as a function of dimensionless income and coincide at high incomes and diverge at low incomes and. Therefore, the density of population at higher incomes, as measured in dimensionless portions of total income, is practically the same in 1990 and 2004. In low-income range, the density of population is relatively higher is 2004. The reason for that is likely not related to increasing income inequality but lower (and varying) accuracy of income measurements at smaller incomes.

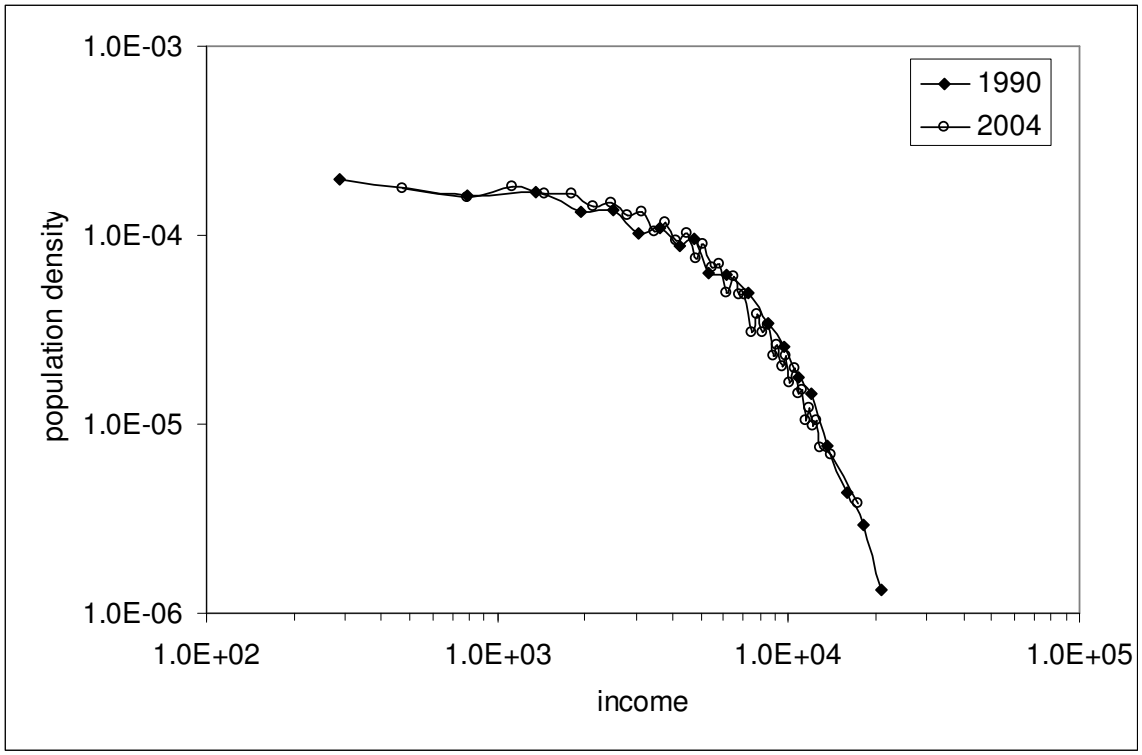

Figure 4. Density of population as a function of scaled income, as reported by the US Census Bureau for 1990 and 2004. The procedure of normalization if the same as in Figure 3. There is no significant discrepancy between these normalized population density distributions. Thus, Gini coefficient does not change with time.

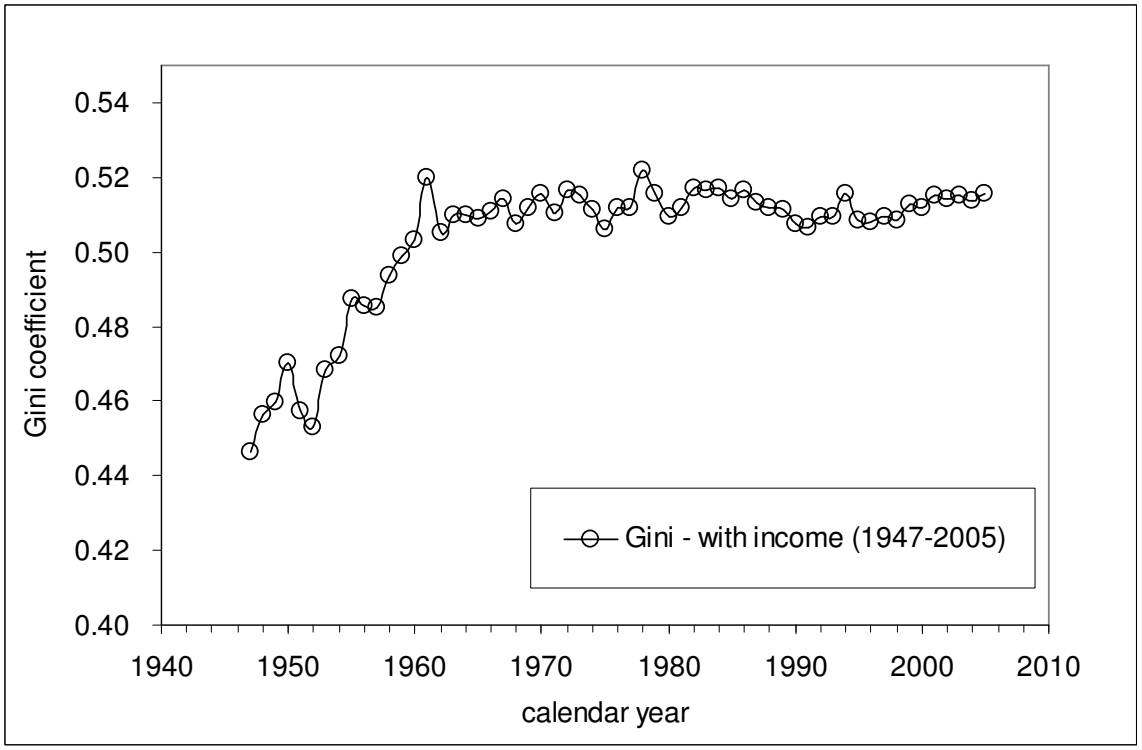

Figure 5. Evolution of empirical Gini coefficient for personal incomes reported by the US Census Bureau between 1947 and 2005. Before 1960, Gini is underestimated due to very low resolution of personal income distribution with only 10 income bins. After 1960, the coverage is better and Gini estimation procedure is reliable. One can conclude that according to the Census Bureau estimates of personal incomes there was no significant change in economic inequality between 1960 and 2005, and also likely before 1960.

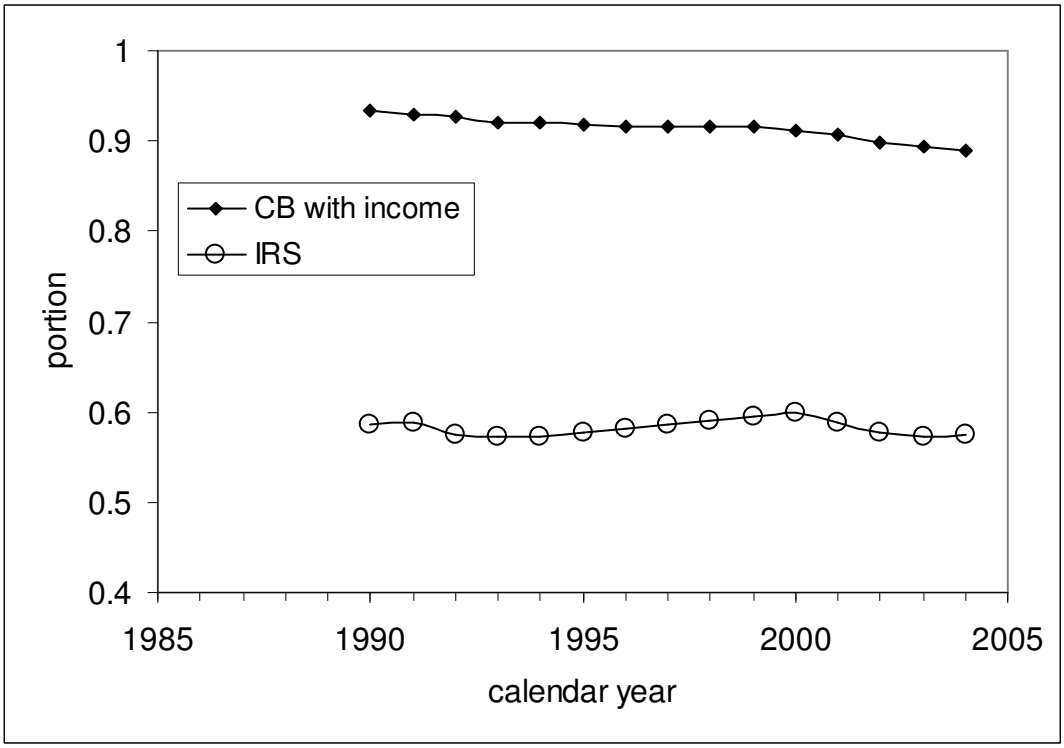

Figure 6. Portion of the working age population with income, as reported by the Census Bureau and IRS. The Census Bureau provides a more reliable definition of income with smaller variations over time and larger portion of working age population with income. Because of very high sensitivity of the number of low income persons to corresponding definition of income the IRS is likely not able to provide a reliable estimate. About 40 percent of working age population is beyond the IRS definition of income.

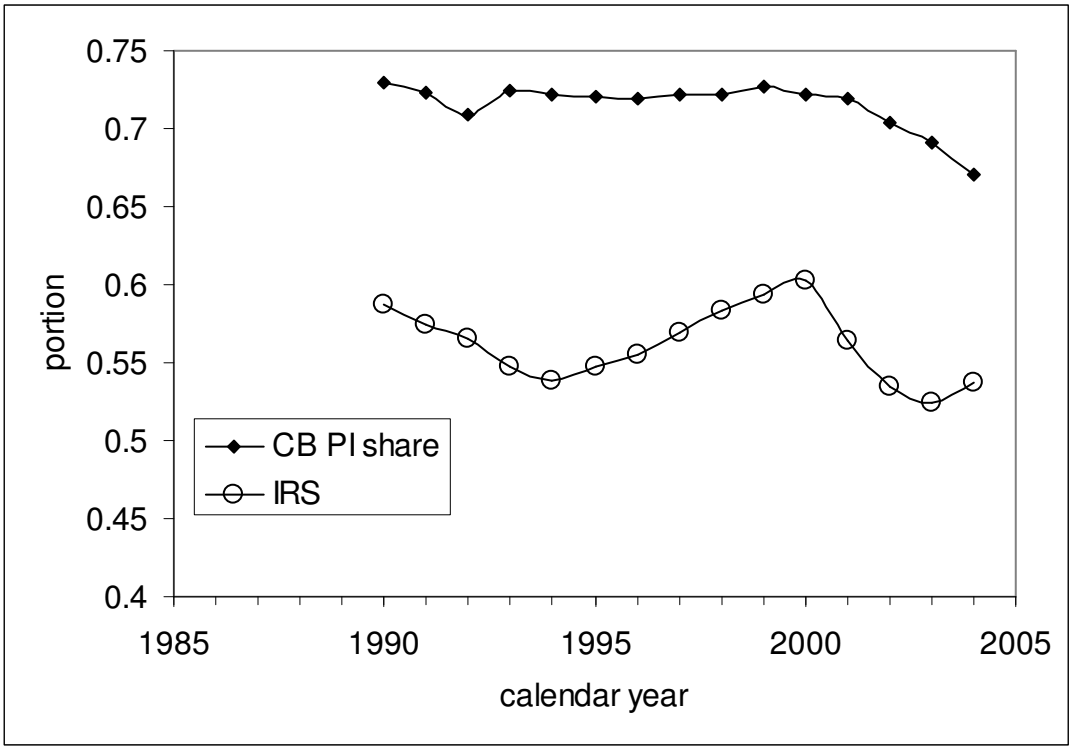

Figure 7. Personal income normalized to real GDP. According to the IRS, only about a half of GDP is transformed in personal income. The BLS reports about 70% of real GDP as personal incomes. Volatility of the IRS is very high.