# The Salamander: A Model of the Right Tail of the Wage Distribution Truncated by Topcoding[1]

**John Angle**
Economic Research Service, 1800 M Street, NW, Washington, DC 20036
jangle@ers.usda.gov

## A Model That Might Facilitate a Lowering of the Minimum Topcodeable Wage and Salary Income

This paper proposes a method to estimate the truncated right tail of the annual wage and salary income distribution using the rest of the distribution.[2] The right tail of the wage and salary income distribution is the distribution of earners over large wage and salary incomes. See figure 1. The metaphor of regenerating a tail that has been cut off suggests the name of this paper's model: the Salamander. The model is in need of a name more compact than "a mixture of gamma probability density functions (pdfs) with parameters quite restricted in a particular way". The U.S. Bureau of the Census masks, or "topcodes", large annual wage and salary incomes in public use microdata samples (PUMS[3]) of the March Current Population Survey (CPS) (Weinberg, Nelson, Roemer, and Welniak, 1999) and other household surveys and censuses to prevent the disclosure of the identity of respondents reporting large incomes. The Federal Committee on Statistical Methodology (FCSM, 1994:62) sees a growing risk of disclosure of the identity of respondents whose data appear in a PUMS particularly those with rare traits, such as large incomes. Angle and Tolbert (1999) discuss the impact of topcoding on the analysis of annual wage and salary income data in the PUMS of the March, CPS. The Salamander was developed to estimate statistics of annual wage and salary income from a PUMS despite topcoding which effectively truncates the right tail of the annual wage and salary income distribution at the minimum topcodeable income.

If the minimum topcodeable income is exactly at the 99$^{th}$ percentile, topcoding only truncates 1% of the distribution, a small proportion of the sample likely to be outliers because of their unusually large incomes. The exclusion of an outlying 1% of the sample from a regression analysis of wage and salary incomes is not an important loss of information in a study of the whole labor force. The loss of information from topcoding incomes does not affect the estimation of any quantile statistic of income smaller than the minimum topcodeable income. However some commonly estimated sample statistics of wage and salary income are substantially influenced by the presence or absence of a few large incomes in the sample and thus by topcoding. Among such statistics are the Gini concentration ratio and the mean of wage and salary income.[4] Given that the distribution of annual wage and salary income is right skewed, the percentage of total

---

[1]Title suggested by an article by Andrew Pollack in the **New York Times** (Late New York Edition), Tuesday, September 24, 2002, pages F1, F4 about the ability of salamanders to regenerate truncated limbs. This project has benefitted from reviews by Linda Atkinson, Robert Gibbs, Charles Hallahan, and Lorin Kusmin of ERS, and Paul Siegel of the Census Bureau. Mistakes and errors of judgment are the author's alone. The views and opinions expressed in this paper do not necessarily reflect the views of the Economic Research Service or the U.S. Department of Agriculture.

[2]See Appendix A on data and methods.

[3]The expression 'public use microdata sample' or PUMS is used by the Census Bureau to denote a sample of individual records from a Decennial Census rather than, for example, the records of its Current Population Survey. This paper uses the term generically.

[4]Since the March 1996 CPS (which reports calendar year 1995 income) it has been possible to estimate the sample mean from a March CPS PUMS because since then the Census Bureau has used the mean of topcoded incomes in several demographic categories rather than the minimum topcodeable income as the topcode. No change was made to the PUMS of earlier March, CPS'.

aggregate income received as incomes equal to or greater than the minimum topcodeable income is perhaps much bigger than 1% and cannot be ignored the way an outlying 1% of the sample can be ignored. The Salamander provides an estimate of the mean of annual wage and salary income. Its estimate of the truncated right tail can substitute for topcoded observations in estimating statistics such as the Gini concentration ratio. The Salamander may work well enough to facilitate a lowering of the minimum topcodeable income below the 99$^{th}$ percentile, if disclosure concerns require better protection of the identities of respondents reporting large incomes.

Three Kinds of Evidence for the Salamander

1) Parsimonious Fit

Three kinds of evidence are offered in this paper to illustrate the validity and/or utility of the Salamander. The first is evidence that the Salamander is a valid model of the whole distribution because it fits March CPS data on the distribution of annual wage and salary income over a period of 41 years (1961-2001) with the estimation of only a few, time-invariant parameters, i.e., parsimoniously. This demonstration is done via a fit of the Salamander without weighting, other than the CPS case weighting required to estimate relative frequencies. When the Salamander is fitted to the entire labor force, the population 16 years of age and older earning at least $1 in a calendar year, it is fit to the distribution, annual wage and salary income conditioned on education and age. If age is restricted to people 25 years of age and older, age need not be used as a variable in the Salamander. Eliminating age makes for a more parsimonious model, one conditioned on education alone. With age restricted to people 25 and over, estimating the Salamander with 41 years of data, i.e., 1961-2001, requires only estimating one parameter for each level of education distinguished. This one parameter is for the fit of the model to the 41 partial distributions[5] corresponding to that level of education, 1961-2001.

2) Algebra of the Salamander Implies Five Distribution Dynamics Seen in Data

The second form of evidence in support of the Salamander's validity as a model of wage and salary income distribution comes from inspection of the Salamander's algebra. This inspection shows that the Salamander is static unless it is driven by change in one or two exogenous variables. These variables are 1) the unconditional mean of wage and salary income in a year, and/or 2) the distribution of education of education in the labor force in a year. The absence of endogenous dynamics makes the Salamander more parsimonious. Change in the unconditional mean of wage and salary income is the more important driver of change in the model for two reasons. This variable is the more important source of change because of the algebra of the Salamander and the fact that empirically it changes more quickly proportionally than the distribution of education in the labor force. All variables in the Salamander are operationalized; they correspond to observed variables in the PUMS of the March CPS. A strong confirmation of the validity of the Salamander as a model of wage and salary income distribution is that its algebra implies five distinctive patterns seen in the dynamics of wage and

[5] 'Partial distribution' is used in the following sense. Where $f(x|y)$ is the distribution of income, x, conditioned on education level, y, and y is a discrete variable, $y_i$, i = 1,2,3,4,5, then $f(x|y_i)$ is the partial distribution of income conditioned on education level i.

salary income distribution in the March CPS wage and salary income data from 1961 through 2001. Although readily demonstrated graphically, some of these patterns are not commented on in the economics literature.

3) Comparison of Salamander Estimates of Mean Wage and Salary Income to Benchmark Estimates

The third form of evidence for the Salamander is crucial for its ability to estimate what people care about in the extreme right tail of the annual wage and salary income distribution. The extreme right tail is not important because of the small fraction of the population it contains. Rather the extreme right tail is important because the sum of the large incomes received in it may be a substantial fraction of the total of income received in the whole population. The functional form of the extreme right tail of the distribution has to be inferred from the total of income received in it or its mean because it is very difficult with the data at hand to estimate this part of the distribution. The largest incomes are suppressed from public use samples by topcoding and few, if any, large household-type surveys are optimized to measure the distribution of the largest incomes. If the Salamander comes close to estimating the unconditional mean of wage and salary income, it, implicitly, comes close to estimating the mean of the extreme right tail of the distribution. There is a benchmark for mean annual wage and salary income. In an internal Census Bureau study to evaluate the adequacy of the estimates of large incomes based on the March CPS, Roemer (2000) established benchmark estimates of mean annual wage and salary income for the years 1990-1996. The source of the benchmark estimates is the the National Income and Product Account (NIPA) estimate of the U.S. Bureau of Economic Analysis (BEA) based on administrative sources of information (U.S. Bureau of Economic Analysis, 2003). Roemer (2000) refers to the NIPA estimates as "benchmarks". Roemer (2000) adjusts the NIPA estimates for 1990-1996 for differences between it and the Census Bureau's CPS definition of annual wage and salary income as well as differences between the Census Bureau's CPS definition of recipient population and the BEA's. Roemer (2000) provides data that allow the comparison of the March CPS sample mean, estimated from the untopcoded March CPS data available in-house at the Census Bureau to the implied NIPA estimate of the mean using Roemer's estimate of the size of the labor force.

Because the population of the comparison is the entire labor force 16+ years of age, producing a comparable Salamander estimate requires estimating the Salamander over the same definition of the labor force. This Salamander estimate requires dealing explicitly with age, expanding the number of model parameters to be estimated. To emphasize that this version of the Salamander is different from the one estimated over the labor force 25 years of age and older and to emphasize that this version of the Salamander is intended to yield good estimates of the far right tail of the distribution truncated by topcoding (in order to yield a good estimate of the mean of wage and salary income), this version of the Salamander is called the "Topcode Salamander". The Topcode Salamander requires the estimation the product k (# of categories of education) and m (# categories of age) parameters. The Topcode Salamander's estimates of the overall (unconditional) mean of wage and salary income in 1990-1996 are then compared to Roemer's (2000) March CPS estimate and adjusted NIPA estimate. The Topcode Salamander estimates of mean annual wage and salary income are based on fits

to March CPS data from 1961 through 2001, not just 1990-1996, and only to the distribution up to $80,000 (in constant 2001 dollars). The Topcode Salamander's 1990-1996 estimates of the mean are made simultaneously with estimates for 34 other years. To demonstrate the robustness of the Salamander against a lowering of the minimum topcodeable income, the Topcode Salamander is re-estimated with income data only up to the unconditional third quartile of income in each of the 41 years of data. Both versions of the Topcode Salamander yield estimates of mean annual wage and salary income in 1990-1996 closer on average to the NIPA benchmarks than Roemer's (2000) estimates based directly on March CPS data.

The Origins of the Salamander

The Salamander was devised as a solution to the problem of inferring the difference between the Gini concentration ratios of nonmetropolitan (nonmetro)[6] annual wage and salary incomes and metro ones. The metro distribution has a heavier right tail and consequently a larger proportion of its incomes are topcoded as a result of a given minimum topcodeable income than is the nonmetro distribution. See figure 1. This differential right censoring of the two distributions interferes with the estimation of the difference between the two distribution's Gini concentration ratios. See Angle and Tolbert (1999) for a discussion of this problem.

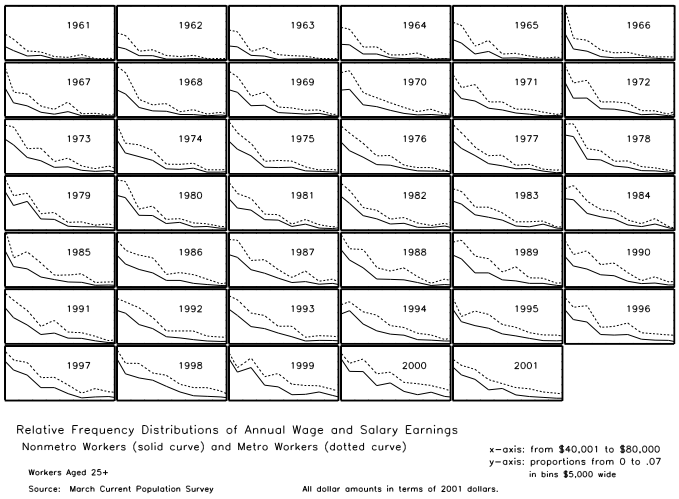


Figure 1: Nonmetro, Metro Right Tails

Figure 1, the right tails of the nonmetro and metro annual wage and salary income distributions from 1961 through 2001, shows that in each year the nonmetro distribution has had a lighter right tail than the metro distribution. Thus the minimum topcodeable income, whatever it is in a given year, is at a higher percentile in the nonmetro than in the metro distribution, putting a larger proportion of large nonmetro incomes at risk of disclosure than metro incomes. Even if a nonmetro income is topcoded, the minimum topcodeable income may be at such a high percentile in the nonmetro distribution that topcoding offers little protection. The higher the percentile of the minimum topcodeable income, the less concealment of larger topcoded incomes among smaller topcoded incomes. Disclosure risk is not hypothetical in the PUMS of the March, CPS. There is a published effort to use machine learning algorithms to perfect a formula to identify which cases in a March CPS have an annual wage and salary income in excess of the minimum topcodeable income (Bauer and

[6] 'Nonmetropolitan' refers to the set of nonmetropolitan counties in the U.S. A nonmetropolitan county is a county not in a Metropolitan Statistical Area (MSA) as defined by the Office of Management and Budget in 1993. MSA's include core counties containing a city of 50,000 or more people or having an urbanized area of 50,000 or more and total area population of at least 100,000. Additional contiguous counties are included in the MSA if they are economically integrated with the core county or counties. The metropolitan status of every county in the U.S. is re-evaluated following the Decennial Census. While there has been a net decline in counties classified as nonmetro over the decades, the definition of nonmetro has remained relatively constant until the 2003 redefinition. See the webpage of the Economic Research Service, "Measuring Rurality" (http://www.ers.usda.gov/briefing/rurality).

Kohavi, 1999). The expected value of most topcoded incomes is not much in excess of the minimum topcodeable income. In a conditional distribution where the minimum topcodeable is at an extremely high percentile, such a formula might defeat topcoding.

Because of the geographic sparseness of the nonmetro population, there is more geographic information per nonmetro case than metro case in a CPS PUMS facilitating the re-identification via matching to a commercial database. The double sparseness of large nonmetro incomes (lighter right tail of wage and salary distribution, more dispersed geographically) exposes the CPS PUMS records of higher income people with a nonmetro residence to a greater likelihood of re-identification via record matching than people with a metro residence earning the identical wage and salary income. Concern to protect this and other vulnerable subsets of the population from disclosure may encourage the U.S. Bureau of the Census to lower the minimum topcodeable income substantially below the 99th percentile. Currently, the minimum topcodeable annual wage and salary income in the March CPS is $150,000. This paper shows that it is possible to model the distribution of annual wage and salary income adequately for at least some purposes with a lower minimum topcodeable income. Disclosure concerns about the PUMS of Federal surveys are given close attention in FCSM (1994).

**The First Kind of Evidence: Parsimonious Fit**

The Salamander is an approximation to the equilibrium distribution of a stochastic process described in Angle (2002b), called the "Inequality Process". A simplified version of the Salamander was used to model how the distribution of annual wage and salary income in the nonmetropolitan U.S. changes shape as a function of change in its mean, and, in particular, how it would change if its mean decreased by 50% (2002a).[7] The Salamander is a

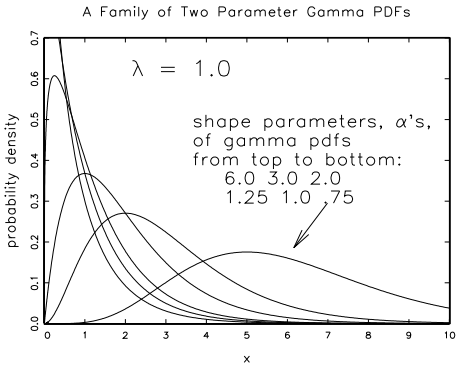


Figure 2: Gamma PDF's

[7] Angle (2002a)'s model differs from this paper's at approximate equation (6) defining $\lambda_{it}$. Instead of

$$\lambda_{it} \approx \frac{(1-\omega_i)\left(\frac{w_{1t}}{\omega_1} + \frac{w_{2t}}{\omega_2} + \ldots + \frac{w_{6t}}{\omega_6}\right)}{\bar{x}_t} > 0 \qquad (6)$$

Angle (2002a), defines a $\lambda_t$ as:

$$\lambda_t \approx \frac{\left(\frac{w_{1t}(1-\omega_1)}{\omega_1} + \frac{w_{2t}(1-\omega_2)}{\omega_2} + \ldots + \frac{w_{6t}(1-\omega_6)}{\omega_6}\right)}{\bar{x}_t} > 0$$

$$\approx \frac{\bar{\alpha}_t}{\bar{x}_t}$$

where $\bar{\alpha}_t$, is the mean shape parameter of the gamma probability density function fit to the partial distributions of the distribution of annual wage and salary income conditioned on education in a particular year. $\lambda_t$ is a simplification of and approximation to (6). The present paper's model including (6) is more faithful to the Inequality Process and fits the data more closely. The advantage of $\lambda_t$ is that it can be presented, as in Angle (2002a) as an

static model of the distribution of wage and salary income, conditioned on education. The Salamander is a mixture. Its components are gamma probability density function (pdf) models, each one of a the partial distribution of this conditional distribution. The mixing weights of the mixture are the proportions of the labor force at each level of education. The shape and scale parameters of each gamma pdf component of the mixture are functions of the Inequality Process parameter for the corresponding level of education. This parameter, $\omega_i$ for the $i^{th}$ level of education is not time-varying. Change in the Salamander is driven by two exogenous variables, a) change in the unconditional mean of wage and salary income, and b) change in the distribution of education in the labor force. Figure 2 displays gamma pdfs with different shape parameters and a fixed scale parameter.

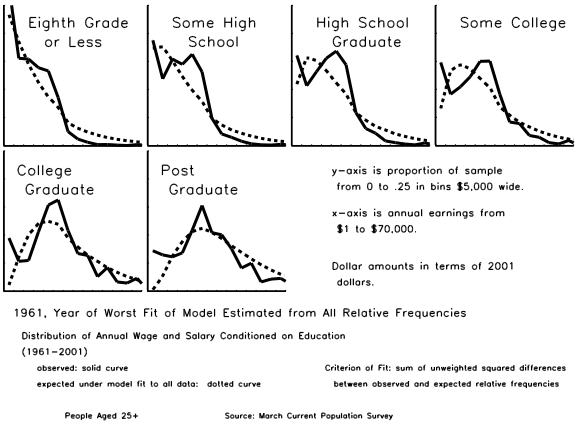


Figure 3: Year of Worst Fit of the Salamander to Conditional Distribution

The gamma probability density function (pdf) has long been used to model income distributions. The gamma pdf is graphed in figure 2 with a fixed scale parameter and a varying shape parameter. Dagum (1977) cites March (1898) as the first published instance. Applications since have been desultory, most appearing in the 1970's, e.g., Peterson and von Foerster (1971), Salem and Mount (1974), Shorrocks (1975), and Boisvert (1977). McDonald and Jensen (1979), assuming the relevance of a two parameter gamma pdf model to empirical income distribution, derive expressions for the statistics of inequality of a gamma pdf in terms of its parameters. Cowell (1977) mentions the gamma pdf as a model of income distribution, if not the best known. The two parameter gamma probability density function (pdf) is:

$$f(x) \equiv \frac{\lambda^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} \quad (1)$$

where:

f(x) ≡ pdf model of the distribution of income

x ≡ income > 0

and:

α ≡ shape parameter of income distribution

arbitrarily constrained gamma probability density function model of income distribution without having to make reference to the Inequality Process and its parameter, $\omega_i$. While Angle (2002a) argues that this simplified model is able to account for the five dynamic patterns in the distribution of annual wage and salary income discussed in the present paper, their explanation in terms of the present paper's algebra is more elegant and the fit of this model to data is closer.

and:

$\lambda \equiv$ scale parameter of distribution of income

The Salamander is not a mixture of unconstrained two parameter gamma pdfs. Such a model is flexible but unparsimonious. For example, with 41 years[8] of data on the distribution of wage and salary income conditioned on 6 levels of education, there are 246 partial distributions. Fitting each one of these by a two parameter gamma pdf, as for example in Salem and Mount (1974), would require 492 parameter estimates. Instead, the Salamander requires just 6 parameter estimates, one parameter estimate, $\omega_i$, for each level of education distinguished in the labor force, to fit the distribution of wage and salary income, conditioned on education, from 1961 through 2001. With these 6 estimated parameters, the model fits 41 years X 6 levels of education = 246 partial distributions, each described by vector of ordered pairs, { income bin mean, relative frequency}. There are 18 income bins from $1 to $5,000 to $85,001 to $90,000, so the model must fit 246 partial distributions X 18 bins = 4,428 relative frequencies (one parameter for each level of education, or 1/41 of a parameter per partial distribution fitted, and 1/738 of a parameter per observation fitted), i.e. parsimoniously. The Salamander is not a flexible functional form. Figure 3 shows the year of the worst fit of the Salamander, 1961, while figure 4 shows the year of the best fit, 1989.

The Salamander model of the unconditional distribution of annual wage and salary income is a mixture of gamma pdfs with constrained parameters. Where $f_{it}(x)$ is the Salamander model of the distribution of wage and salary income to people at the $i^{th}$ level of education at time t, the Salamander model of the unconditional distribution of wage and salary income, $f_t(x)$, is:

$$f_t(x) = w_{1t} f_{1t}(x) + ... + w_{it} f_{it}(x) + ... + w_{6t} f_{6t}(x) \qquad (3)$$

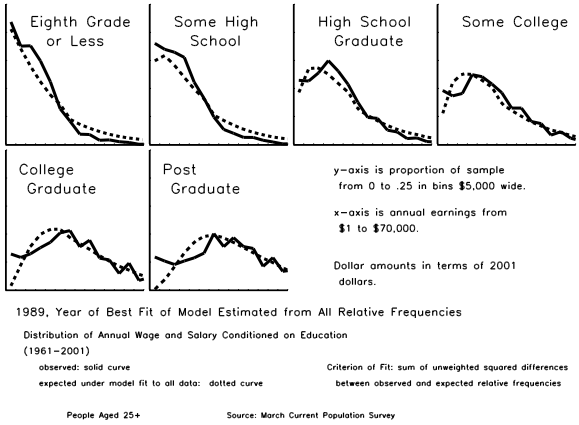


Figure 4: Year of Best Fit of the Salamander to Conditional Distribution

where $w_{it}$, the mixing weight, is the proportion of the labor force at the $i^{th}$ level of education in a particular year. The Salamander component model of each partial distribution of the

[8] Since no data on education are available from the 1963 March Current Population, the distribution of 1962 wage and salary income conditioned on education is estimated by interpolation from 1961 and 1963 data. There are thus 40 years of data over 41 years. See Appendix A.

distribution of wage and salary income conditioned on education is a gamma pdf with constrained parameters:

$$f_{it}(x) \equiv \frac{\lambda_{it}^{\alpha_i}}{\Gamma(\alpha_i)} x^{\alpha_i - 1} e^{-\lambda_{it} x} \quad (4)$$

where:

$f_{it}(x)$ ≡ pdf model of the distribution at level i of education at time t

x ≡ income > 0

and:

$\alpha_i$ ≡ shape parameter of income distribution of people at level i of education

$$> 0$$

$$\approx \frac{1 - \omega_i}{\omega_i} \quad (5)$$

and:

$\lambda_{it}$ ≡ scale parameter of distribution of income of people at level i of education at time t

$$> 0$$

$$\approx \frac{(1-\omega_i)\left(\frac{w_{1t}}{\omega_1} + \frac{w_{2t}}{\omega_2} + \ldots + \frac{w_{6t}}{\omega_6}\right)}{\bar{x}_t} \quad (6)$$

where:

$\omega_i$ ≡ a parameter of the Inequality Process (Angle, 2002b) the stochastic interacting particle system; the Salamander approximates the equilibrium distribution of the Inequality Process.

$\bar{x}_t$ ≡ unconditional mean of income at time t

$w_{it}$ ≡ proportion of population at level i of education at time t

and $\bar{x}_t$ and the $w_{it}$'s are exogenous and the sole source of change. Since the $w_{it}$'s vary more slowly proportionally than $\bar{x}_t$ and because of the algebra of (6),

$$\Delta\bar{x}_t \equiv \bar{x}_t - \bar{x}_{(t-1)}$$

is the main source of change from year to year and it comes via change in the scale parameter, $\lambda_{it}$.

$\lambda_{it}$ cannot be estimated unless an estimate of $\bar{x}_t$ is available. There is a way to estimate $\bar{x}_t$ based on sample wage and salary incomes despite topcoding. The conditional medians of wage and salary income can be readily estimated from the public use sample, unaffected by topcoding. The sample median at each level of education, $M_{it}$ , is an excellent estimate of the corresponding population level statistic. An estimate of $\bar{x}_t$ is implied by the search over $\Omega$, the $\omega_i$'s, the parameter vector, in the fitting of the model. If a benchmark for $\bar{x}_t$ exists it can be supplied directly to the Salamander.

Obtaining an estimate of $\bar{x}_t$ in terms of the $M_{it}$'s and the $w_{it}$'s, under the model, requires observing that the median of a gamma pdf is approximately (Salem and Mount, 1974):

$$M_{it} \approx \frac{(3\alpha_i - 1)}{3\lambda_{it}} \tag{7}$$

and the mean of a gamma pdf is the ratio of the shape parameter to the scale parameter. So the Salamander's estimate of the conditional mean is:

$$\bar{x}_{it} \approx \frac{\alpha_i}{\lambda_{it}} \tag{8}$$

$\bar{X}_t$ is estimated, under the model, as:

$$\bar{X}_t \approx \sum_{i=1}^{I} w_{it} M_{it} \left[ \frac{(1 - \omega_i)}{\left(1 - \frac{4}{3}\omega_i\right)} \right] \tag{9}$$

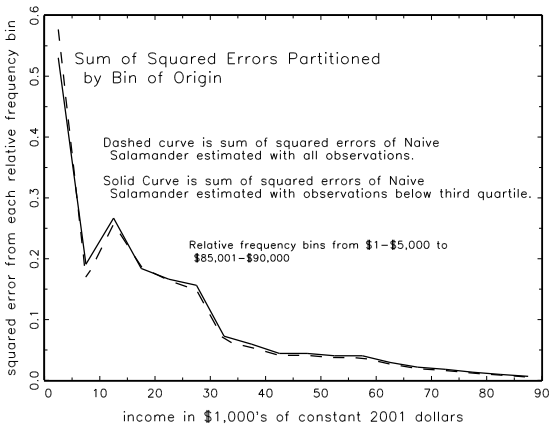


Figure 5: Salamander's Sum of Squared Errors Partitioned by Bin of Origin

**Fit and Estimation of the Salamander**

The six $\omega_i$'s, one for each level of education distinguished, are given in Table 1, first as estimated using all the relative frequencies and then as estimated using the relative frequencies smaller than the $Q_{3i}$, the third quartile of each of the partial distributions of the distribution of annual wage and salary income, 1961-2001, conditioned on level of education. The estimates and fits of Table 1 are done without weighting, other than the individual case weighting involved in estimating the relative frequencies.

Table 1

| Highest Level of Education | $\omega_i$ Estimated by Fitting Salamander to All Relative Frequencies | Bootstrapped Standard Error of $\omega_i$ Estimated by Fitting Salamander to All Relative Frequencies (27 re-samples) | $\omega_i$ Estimated by Fitting Salamander to Relative Frequencies below $Q_{3i}$ | Bootstrapped Standard Error of $\omega_i$ Estimated by Fitting Salamander to Relative Frequencies below $Q_{3i}$ (27 re-samples) |
|---|---|---|---|---|
| Eighth Grade or Less | .4753233 | .0007434 | .4820085 | .0010150 |
| Some High School | .4377520 | .0008516 | .4448130 | .0009233 |
| High School Graduate | .3805134 | .0004859 | .3958540 | .0008192 |
| Some College | .3423497 | .0012267 | .3594217 | .0009410 |
| College Graduate | .2642188 | .0006449 | .2713351 | .0014287 |
| Post Graduate Education | .2263275 | .0008740 | .2292268 | .0009688 |

Note that while there is a systematic difference between the two sets of parameters, both sets are close. The graphs of the fitted models, for example in figures 3 and 4, are indistinguishable from each other. The inverse association between $\omega_i$ and level of education is stable and widespread throughout the industrialized world (Angle, 1993). Note that the fits of the partial distributions in the year worst fit by both estimates of the Salamander, 1961, is fairly good. See figure 3 and compare with figure 4, which shows the best fit of the Salamander, in 1989.

Table 2 shows that the fit of the Salamander estimated with income observations below $Q_{3i}$ is not as close as that of the Salamander estimated in terms of all the data, but the difference is small. Table 2 shows the close fit of the Salamander to 246 distributions, observations made over 41 years of change. Most of its sum of squared errors come from bins in the left tail of the income distribution which makes the left tail the most important part of the distribution to be fitted. See figure 5. The closest possible fit of a gamma pdf model to these 246 distributions is the fit of 246 unconstrained two parameter gamma pdfs, one to each of the 246 partial distributions. This closest possible gamma pdf fit requires 492 degrees of freedom. Yet, its fit is not substantially better than the Salamander's with its 6 degrees of freedom. The sum of squared errors of the unconstrained two parameter gamma pdf fits is 1.422 (See Table 2). The Salamander's sum of squared errors is only 31.4% larger with 486 fewer parameters to estimate. The sum of squared errors of the Salamander estimated with data up to $Q_{3i}$ is only 32.7% larger than the closest possible gamma pdf fit. When fit is measured more robustly in terms of the sum of absolute deviations, the comparisons are even starker. The Salamander's sum of absolute deviations is only 9.2% larger than that of the best possible gamma pdf model, while the Salamander estimated with data up to $Q_{3i}$ has a sum of absolute deviations that is only 12.1% larger than that of the closest possible fitting gamma pdf model with 1/82 the number of parameters of the closest possible fitting gamma pdf model.

Table 2

| Criterion of Fit | Fit of Salamander to All Relative Frequencies up to $90,000 | Fit of Salamander to All Relative Frequencies up to $90,000 with Parameters Estimated from Relative Frequencies only up to $Q_{3i}$ | Fit of 246 unconstrained 2 parameter gamma pdfs to All Relative Frequencies up to $90,000 |
|---|---|---|---|
| Sum of Squared Errors | 1.868667 | 1.886785 | 1.422116 |
| Mean Squared Error Per Bin | .0003798 | .0003835 | .0002890 |
| Sum of Absolute Errors | 64.98238 | 66.69445 | 59.50923 |
| Mean Absolute Error Per Bin | .0132078 | .0135557 | .0120954 |
| Correlation Between Observed and Expected | .9289116 | .9290207 | .9463818 |

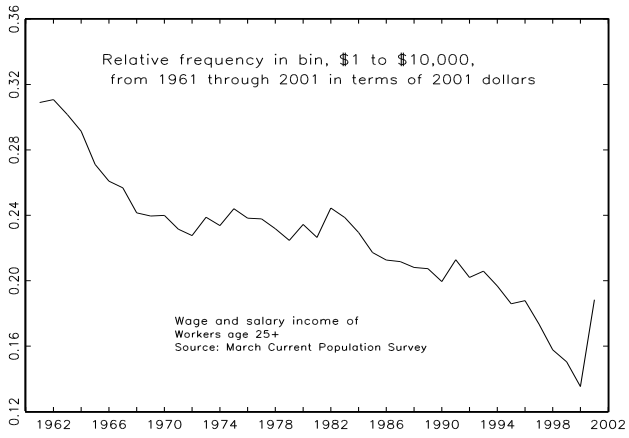


Figure 6: Relative Frequency in Bin $1-$10,000

**The Second Kind of Evidence: The Salamander Implies Five Distribution Dynamics Seen in Data**

The Salamander accounts for five dynamic patterns seen in the distribution of wage and salary income via the algebra of the partial derivative of (4) with respect to the unconditional mean, i.e., $\partial f_{it}(x) / \partial \bar{x}_t$. These five dynamic patterns are:

A. Left tail frequencies become smaller and right tail frequencies become greater when the unconditional mean increases;

B. Relative frequencies of left and right tails are inversely correlated (as $\bar{x}_t$ increases, left tail thins and right tail thickens);

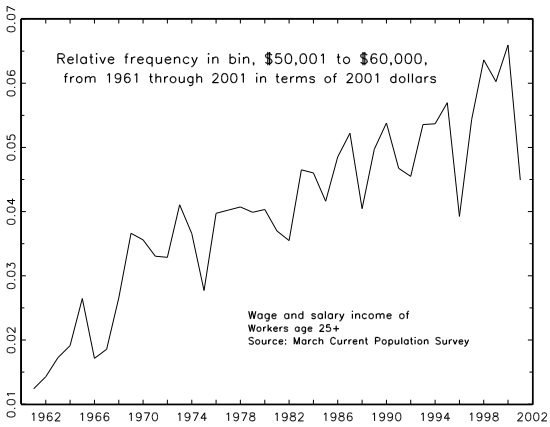


Figure 7: Relative Frequencies in Bin $50,001-$60,000

C. The greatest mean proportional change in the distribution occurs in the relative frequencies of the largest wage and salary incomes;

D. There is quasi-symmetric change in the relative frequencies of the left and right tails at incomes equidistant from mean of wage and salary income; these changes are quasi-symmetric in the sense that they are nearly proportionally equal although opposite in sign;

E. Stochastic variability in relative frequency is proportional to relative frequency;

via the partial derivative of the model of the relative frequency of wage and salary income in a bin of the distribution of earners at the $i^{th}$ level of education with respect to to $\bar{x}_t$, the principal exogenous shock and source of change in model. This partial is:

$$\frac{\partial f_{it}(x)}{\partial \bar{x}_t} = f_{it}(x) \cdot \lambda_{it} \cdot \left( \frac{x - \bar{x}_{it}}{\bar{x}_t} \right) \tag{10}$$

where:

$$\lambda_{it} \approx \frac{(1-\omega_i)\left(\frac{w_{1t}}{\omega_1} + \frac{w_{2t}}{\omega_2} + \ldots + \frac{w_{It}}{\omega_I}\right)}{\bar{x}_t}$$

$$\bar{x}_{it} \approx \frac{\alpha_i}{\lambda_{it}}$$

See Appendix B for the derivation of (10).

Dynamic Pattern A: Fewer Small Incomes, More Large Incomes, as Unconditional Mean Increases

Equation (10) states that the rate of change in the relative frequency of the bin whose mean is x in the distribution of wage and salary income to earners at the $i^{th}$ level of education is proportional to the size of the relative frequency, to that partial distribution's scale parameter, and to the ratio of the signed difference between x and the conditional mean to the unconditional mean. It is this latter term:

$$\left(\frac{x-\bar{x}_{it}}{\bar{x}_t}\right) \tag{11}$$

that explains Dynamic Pattern A, the way the left and right tails of the distribution of wage and salary income flux when the unconditional mean of wage and salary income, $\bar{x}_t$, increases. The first two terms of (10) are always positive. The third term (11) is negative to the left of the conditional mean, $\bar{x}_{it}$, positive to the right. Consequently when $\bar{x}_t$ increases, as it did substantially in the period 1961-2001, the partial derivative of $f_{it}(x)$ with respect to $\bar{x}_t$ increases in the right tail and decreases in the left tail of $f_{it}(x)$, the model. Figures 6 and 7 show that the model mimics the empirical left and right tails of the wage and salary income distribution in this respect. Relative frequencies in the far left tail of the distribution fell in this period as the mean of wages and salaries rose, while relative frequencies in the right tail, represented by the bin, $50,001 to $60,000 rose.

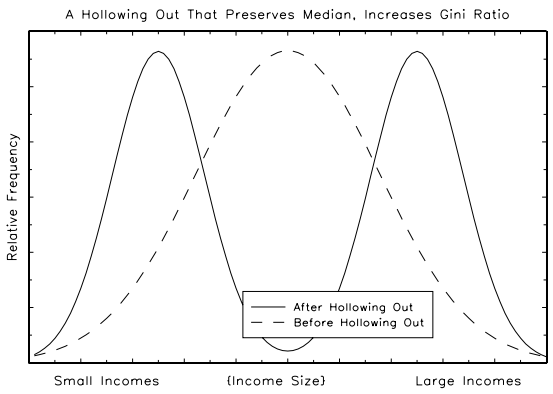


Figure 8: a mis-impression of trend in the wage distribution in the economics literature

This empirical high inverse correlation between the relative

frequencies in the left and right tail bins (figures 6 and 7) is perhaps puzzling because it is unanticipated in the economics literature. Levy and Murnane (1992: 1349) sum up a widely held explanation for the trends of a stagnant median of real earned income in the 1980's and an increasing Gini concentration ratio in a review article in the **Journal of Economic Literature**:

> "A careful reader of these and other articles ... would have emerged with several conclusions. First, the earnings distribution for year-round, full-time male workers was "hollowing out" in the sense implied by a shrinking middle class: The middle of the distribution was declining while the upper and lower tails were growing ...."

i.e., that a positive correlation exists between the relative frequencies of the left and right tails of the distribution as both thicken simultaneously rather than the large negative correlation that actually exists. The "hollowing out" thesis was advanced to account for the time series of two statistics of the wage and salary incomes: a median that showed no substantial increase in the 1980's and at times declined, and a Gini concentration ratio which increased. Levy and Murnane (1992: 1339ff) summarize this literature on the "hollowing out" of the wage and salary income distribution. Figure 8 caricatures such a hollowing out which would leave the median unchanged but greatly increase the Gini concentration ratio.

While unexplained and unanticipated by economic theory, in a general way dynamic patterns A through E do not conflict with the view that a "rising tide lifts all boats" (Danziger and Gottschalk, 1986), that all workers have a community of interest in a business expansion that raises the mean of wage and salary income regardless of the size of their particular wage and salary income. The fall in the left tail and the rise in the right tail indicate that both small and large wage and salary incomes become larger for those continuing in the labor force when the mean increases. But the saying "a rising tide lifts all boats" implies that small and large incomes rise an equal amount. In distributional terms that implies a rigid translation of the distribution to the right. Neither such a translation nor a "hollowing out" occurs in the empirical distribution. These mis-impressions are interpretations of what happened to the distribution based on time-series in statistics of the distribution.

Dynamic Pattern B: Inverse Correlation between Relative Frequencies of Left and Right Tails

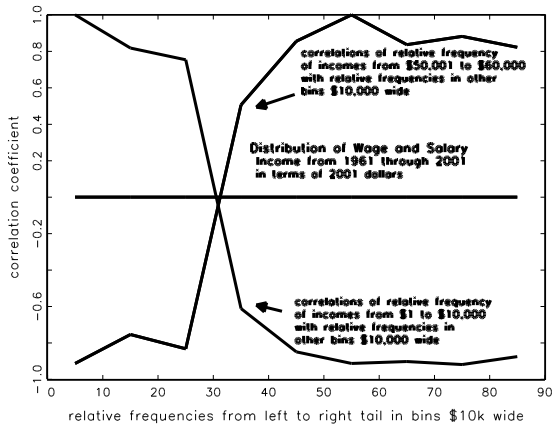


Figure 9: Correlations of relative frequencies in two bins approximately equidistant from the mean in the left and right tail with other relative frequencies

Dynamic Pattern B generalizes Dynamic Pattern A showing the correlation structure over time between the relative frequencies of figures 6 and 7 and the relative frequencies of other bins. Figure 9 shows that as implied by term (11) changes in the relative frequencies in the left and right tail are inversely correlated and have a near zero correlation with relative frequencies bins in the vicinity of the mean. Also as implied by term (11) positive correlations of relative frequencies within the left and right tail attenuate with distance between the bin means. Figure 9 shows the correlation structure of the unconditional distribution, while (10) refers to a partial distribution. However, the relationship aggregates up to the unconditional distribution in a straight-forward way.

Dynamic Pattern C: Greatest Mean Proportional Change in the Distribution Occurs in the Relative Frequencies of the Largest Wage and Salary Incomes in the Distribution

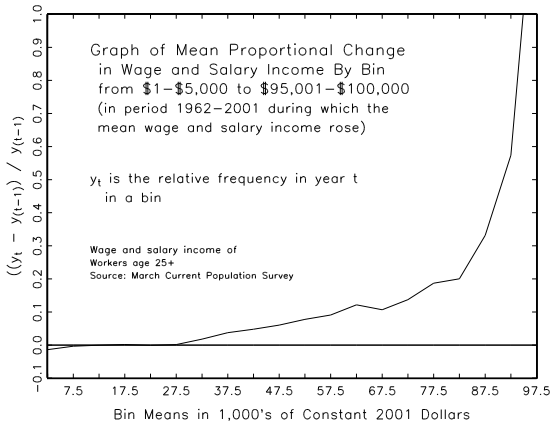


Figure 10: Mean proportional growth of relative frequencyis exponentially greater the larger the income with growth in unconditional mean of income.

Figure 10 shows the mean proportional change of relative frequencies by bin over the period 1961 - 2001, a time period in which the unconditional mean of wage and salary income rose. Term (11) in (10) implies that, when the unconditional mean increases, the biggest proportional change in any relative frequency is in the relative frequency of the largest wage and salary income. This result follows because the difference $(x - \overline{x}_{it})$ is greatest for this bin. In the Salamander, change in the far left tail can never be as large proportionally in absolute value as change in the far right tail, according to term (11), because the distribution is bounded at zero and right skewed and the numerator of term (11) will always have a larger value for the largest income observed than the smallest. Proportional change is defined here as:

$$\left( \frac{y_t - y_{t-1}}{y_{t-1}} \right) \qquad (12)$$

where $y_t$ is the relative frequency in a bin in year t. Expression (12), calculated from empirical relative frequencies, is analogous to the LHS (left hand side) of (13) whose RHS (right hand side) is, at any one time period, a constant multiplied by $(x - \bar{x}_t)$ for increasing x, income.

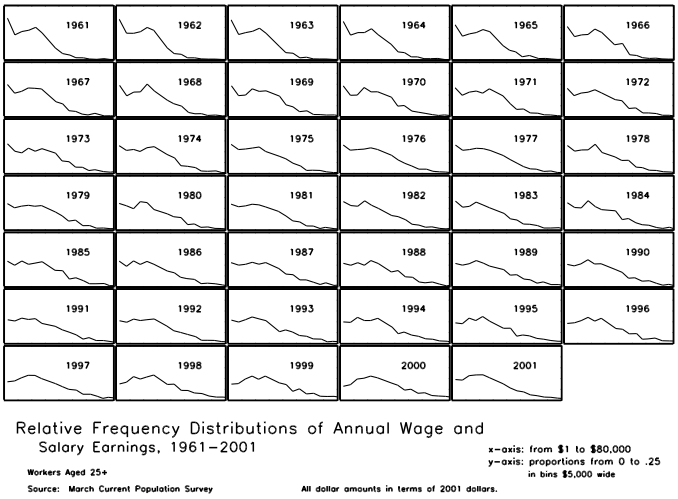


Figure 11: Relative Frequency Distribution of Annual Wage and Salary Income, 1961-2001

$$\frac{\left( \frac{\partial f_{it}(x)}{\partial \bar{x}_t} \right)}{f_{it}(x)} = \lambda_{it} \left( \frac{x - \bar{x}_{it}}{\bar{x}_t} \right) \qquad (13)$$

The thickening of the right tail of the wage and salary income distribution in recent decades, discussed by Levy and Murnane (1992) and in the sociology literature (e.g., Morris, et al., 1994), where it is labeled "polarization" is empirically what has been happening to the U.S. distribution of wage and salary income from 1961-2001 as the unconditional mean increased. No substantial split of the population of wage and salary income recipients occured in the years 1961-2001 when the right tail thickens, because the left tail thins as workers move up (in annual wage and salary income) into the central mass of the distribution. So the thickening of the right tail, which means that more people are receiving larger wage incomes, does not split the distribution into haves and have-nots. The thickening of the right tail as the unconditional mean increases is consistent in the Salamander with the $\alpha_i$, the shape parameter, of the wage and salary distribution of the $i^{th}$ education level remaining constant. The Salamander explains how the right tail of each partial distribution of the conditional distribution, wage and salary income conditioned on education, thickens with growth in the unconditional mean without the creation of a distribution like that of figure 8. As shown in figure 11, the left tail of the distribution thinned, the right tail thickened, and the mass of the distribution shifted rightwards in the period 1961-2001, i.e., fewer very small incomes, more very large incomes, and a general raising of most people's incomes. Figure 11 shows no "hollowing out," no substantial bi-modality, no polarization. Figure 11 differs from figure 8.

Dynamic Pattern D: Quasi-Symmetric Change in Relative Frequencies of Left and Right Tails at Incomes Equidistant From Mean of Wage and Salary Income

There is an even more detailed confirmation of the Salamander's algebra: Dynamic Pattern D, quasi-symmetric change in the tails of the unconditional distribution, after standardization and sign reversal of one, as illustrated in figure 12. Figure 12 graphs the relative frequencies of the unconditional distribution of annual wage and salary income in bins $1-$10,000 and $50,001-$60,000 from 1961 through 2001 after these relative frequencies have been standardized and the signs of the standardized relative frequencies of bin $1-$10,000 have been reversed. The two sets of transformed relative frequencies nearly overlap because the midpoints of their bins are approximately equidistant from the grand mean of unconditional annual wage and salary income 1961-2001: $30,839. Both mid-points are, approximately, $25,000 away. Expression (11) and equation (10) are given for the distribution conditioned on education. However, their implications aggregate to implications for the unconditional distribution in a straightforward way. The conditional mean, $\overline{x}_{it}$, is a scalar constant multiplied by the unconditional mean, $\overline{x}_{t}$. See (10). Equation (10) implies that change differs between relative frequencies equidistant from the conditional mean as the unconditional mean increases only in sign and magnitude. The left tail relative frequencies decrease as the unconditional mean increases while the right tail relative frequencies increase, i.e., they have different signs. Since left tail relative frequencies are larger than right tail relative frequencies, change will be larger in absolute terms in the left tail than the right tail of bins equidistant from the conditional mean. Standardization of left and right tail relative frequencies cancels out the effect of magnitude of the relative frequencies on change, as explained in Appendix C. Consequently, equation (10) implies that the graphs of relative frequencies thus transformed from bins approximately equidistant from the mean should, approximately, be equal. Figure 12 shows that this implication holds empirically. Figure 12 shows that standardization of the relative frequencies of figures 6 and 7 with the sign reversal of one results in a time series that overlaps very closely and is correlated .91, indicating near statistical equivalence.

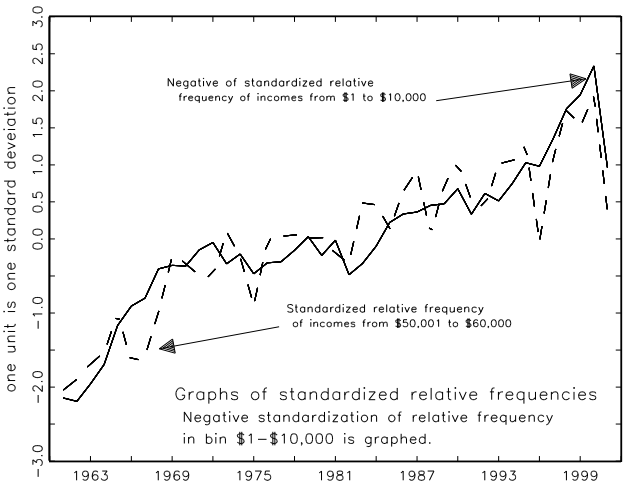


Figure 12: These Relative Frequencies in Left and Right Tail Almost Overlap after Standardization and Sign Reversal of One

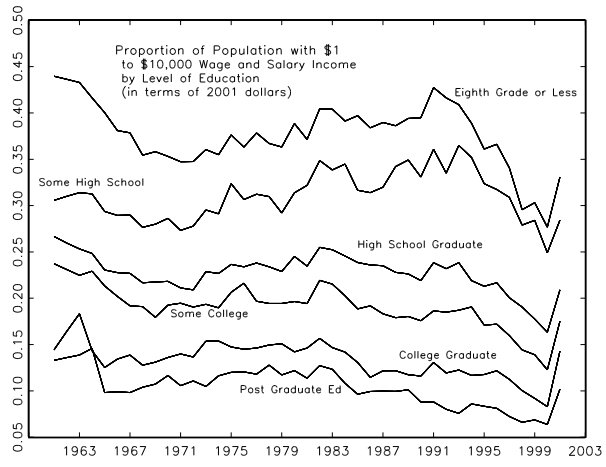


Figure 13: Conditional Relative Frequencies at a Point in Left Tail by Level of Education

Dynamic Pattern E: Stochastic Variability in Relative Frequency Proportional to Relative Frequency

Equation (10) implies that the Salamander's partial derivative of the relative frequency at x, $f_{it}(x)$, with respect to the exogenous stochastic variable $\overline{x}_t$, varies in absolute value with $f_{it}(x)$ itself. Stochastic shocks introduced by the exogenous variables, $\overline{x}_t$, the unconditional mean of wage and salary income, and the $w_{it}$'s, the distribution of the labor force over levels of education. Figures 13 and 14 show graphically that the larger the relative frequency at a given income in left and right tails, the greater the relative frequency, the noisier it is, as implied by (10). Means and standard deviations in figure 13 are correlated .892; in figure 14 that correlation is .988, as implied by the Salamander.

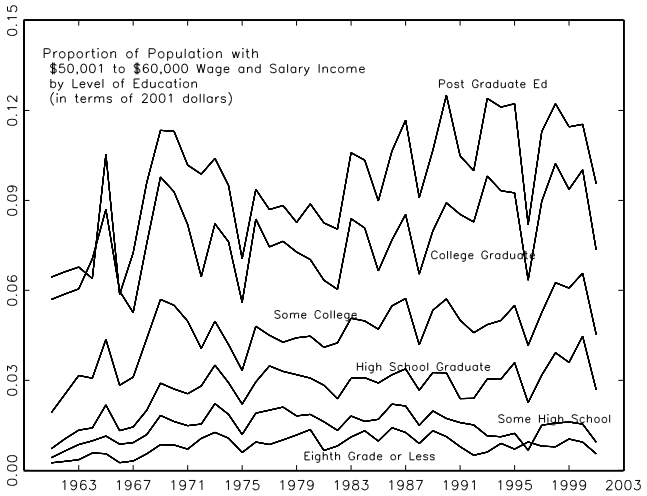


Figure 14: Conditional Relative Frequencies at a Point in Right Tail by Level of Education

**The Third Kind of Evidence: The Topcode Salamander's Estimate of the Unconditional Mean vs. Roemer's (2000) Estimate from Untopcoded March CPS' and the Comparable NIPA Estimate**

It is easy to discard 1% or fewer of a sample from a regression analysis that seeks to generalize from the sample as a whole, particularly when the discarded 1% are outlying observations. Certainly incomes at or above the 99$^{th}$ percentile are unusually large incomes. They are drawn from a skewed distribution of incomes. The skew of this distribution, when the dependent variable of a regression is income, is often responded to by transforming the distribution of the dependent variable by using its logarithm as the dependent variable of the regression. One might argue then that as far as regressions seeking to generalize to relationships between income and other variables in the population as a whole, the topcoding of the largest 1% or fewer incomes is hardly a problem. However, there are statistics of income that are commonly estimated which are sensitively dependent on the far right tail of the distribution of incomes and thus the largest incomes in a sample of incomes. The loss of information about the largest incomes occasioned by topcoding biases estimates of these variables. A statistic as commonly estimated as the mean of annual wage and salary income is among such statistics. Beginning with the March 1996 CPS the Census Bureau removed the obstacle of estimating the unconditional mean of a March CPS sample of annual wage and salary income by using the mean of the topcoded incomes in several demographic categories as the topcode for annual wage and salary incomes larger than the minimum topcodeable income of PUMS records corresponding to that demographic category. This procedure also allows unbiased estimation of the mean of incomes in excess of the minimum topcodeable income of cases in the specified demographic categories. Standard errors of estimates of the mean are biased downward in the cases of the unconditional mean and the means in the specified demographic categories, upward in all other cases.

Prior to 1996 the topcode used in the March CPS had been the minimum topcodeable income. Topcodes in the PUMS of March CPS' prior to 1996 were never revised. Thus estimating mean annual wage and salary income from March CPS' prior to 1996 required metadata about the data lost through topcoding. Traditionally the source of the metadata is the presumption that the truncated right tail is distributed as a Pareto probability density function (pdf). See Henson, 1967, Pareto, 1897; Parker and Fenwick, 1983; Shryock and Siegel, 1973; and Spiers, 1977. The Salamander is intended to fill the same need. No comparison is made between the Salamander and the Pareto pdf because the premise of the Pareto pdf is that there is a constant ratio between the point of right truncation of the distribution and the mean of incomes to the right of that point. This premise is inaccurate (Angle and Tolbert, 1999) in the case of March CPS data on annual wage and salary income as can easily be established with March CPS' from the 1960's when the minimum topcodeable income was so high that almost no incomes were topcoded in the PUMS of those CPS'. Since the Census Bureau itself (Roemer, 2000:17-21) has raised questions about the validity and reliability of the March CPS as a measuring instrument for large incomes, fitting the distribution of large incomes in March CPS' from the 1960's is not necessarily an adequate test of the Topcode Salamander, the version of the Salamander specialized for yielding good estimates of the right tail of the distribution truncated by topcoding as evidenced by its yielding an accurate estimate of the mean of annual wage and salary income. Roemer (2000), "Assessing the Quality of the March Current Population Survey and the Survey of Income and Program Participation Income Estimates, 1990-1996", is the report of the Census Bureau's own internal evaluation of the adequacy of the income data collected by the March CPS.

Roemer (2000) benchmarks the adequacy of annual wage and salary income collected by the March CPS against the total aggregate of annual wage and salary income of the National Income and Product Accounts (NIPA), compiled from administrative sources of information by the U.S. Bureau of Economic Analysis (2003). Roemer (2000) uses March CPS case weighting to find the implied national total aggregate of annual wage and salary income in the CPS. He adjusts the NIPA benchmark for differences between its definition of annual wage and salary income and that of the March CPS and for differences in the population referred to by each definition. He then compares the estimate of the total aggregate of annual wage and salary income in the March CPS' of 1991 through 1997 (covering annual income 1990-1996) with the adjusted NIPA benchmark. He also provides the March CPS' estimate of the recipient population, the labor force. Dividing his March CPS total aggregate annual wage and salary income estimate by his estimate of the size of the recipient population gives his implied estimate of the mean, both the March CPS mean and the comparable NIPA mean. These are graphed in figure 15, after having both been converted from current nominal dollars to constant 2001 dollars. Roemer's estimate of the mean varies

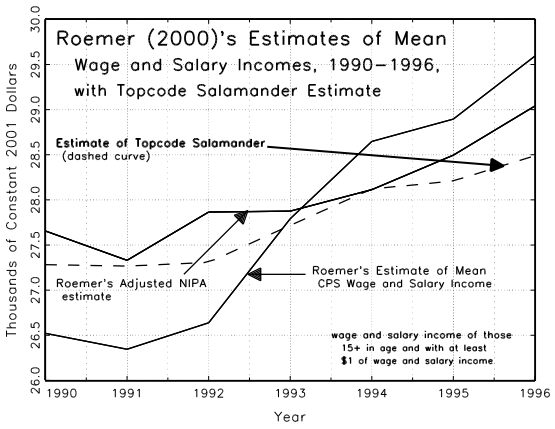


Figure 15: Roemer's estimate of mean March CPS wage income and comparable NIPA estimate vs. estimate of Topcode Salamander

from the NIPA estimate by an average $702 a year (in absolute value).

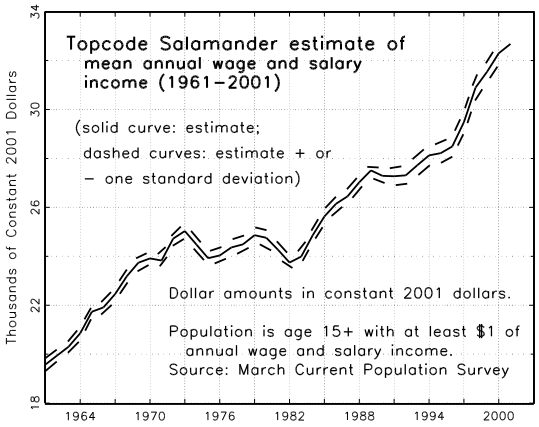


Figure 16

The Topcode Salamander provides via (9) an estimate of the unconditional mean of annual wage and salary income in the population on which it is estimated. The Topcode Salamander's implied mean is graphed in figure 15. The estimate graphed in figure 15 is done in a different way from the earlier fit of the Salamander and that is why it is called the Topcode Salamander. The estimation technique of the Topcode Salamander is described below. It is called the 'Topcode Salamander' to emphasize that it is intended to replace data on large incomes lost to topcoding. The technique used to estimate the Topcode Salamander is described below. The Topcode Salamander estimate is close to Roemer's NIPA estimate, closer in fact than Roemer's estimates based on untopcoded March CPS samples. The absolute difference between the Topcode Salamander's estimates and Roemer's CPS estimates averages $285 a year in absolute value. It should be noted that these Topcode Salamander estimates are not specialized for the years 1990-1996 in any way. Rather they come from a fit of the Salamander to all available March CPS data on annual wage and salary income, 1961-2001. This fit implies 41 estimates of the unconditional mean of annual wage and salary income in each of the 41 years via (9). These are the estimates graphed in figure 15 along with its standard error of estimate based on 50 bootstrapped samples. This bootstrapping samples the March CPS 50 times, estimates relative frequencies 50 times, and then runs the stochastic search algorithm that estimates the parameter vector 50 times. How could the Topcode Salamander estimate be closer to Roemer's benchmark than Roemer's estimate based on untopcoded income data collected by the March CPS? One reason this result is plausible is Roemer's (2000: 17-21) list of reasons why the March CPS is a questionable instrument for measuring large incomes.

Roemer (2000), the Census Bureau's evaluation of the March CPS as an instrument for measuring income, expresses doubt about the reliability of the March CPS' as an instrument to measure large annual wage and salary incomes. Roemer (2000) recounts anecdotes of the operational idiosyncracies of obtaining reports of large incomes with the March CPS (Roemer, 2000: 17-21). He points out that in March 1994 (collection of 1993 annual income data) the Census Bureau began computer assisted interviewing, an event which boosted reports of large incomes much more than smaller incomes, raising questions about what had been held back earlier, why computer assisted interviewing elicited reports of larger incomes, and how extensive under-reporting of large incomes remained. In March 1995, the maximum annual wage and salary income that the questionnaire allowed more than quadrupled, also boosting reports of large incomes. And, comparing reports of annual wage and salary income in the March CPS to Federal individual tax returns, Roemer (2000: 21) finds that "...this correspondence seems to worsen quite suddenly at the high end.", an indication of under-reporting, perhaps severe, of large annual wage and salary incomes to the March, CPS interviewers.

The Topcode Salamander's estimate of mean annual wage and salary income is remarkably robust against the loss of information about large incomes in the March CPS, i.e., robust against a very substantial lowering of the minimum topcodeable income. Figure 17 shows that when the Topcode Salamander is re-estimated using only observations on incomes at or below the 3rd quartile (75th percentile) of annual wage and salary income in a year, the resulting estimate of mean annual wage and salary income is close to its estimate based on the all incomes up to $80,000 (in constant 2001), i.e., at a much higher percentile than the 75th. In the years 1990-1996, the mean absolute difference between the Topcode Salamander estimate of mean annual wage and salary income when income observations are limited to those at or below the 3rd quartile in each year and Roemer's adjusted NIPA estimate is only $180 a year, substantially less than that of the first Topcode Salamander estimate based on incomes up to $80,000 in constant 2001 dollars. See figure 18. Neither estimate is specialized for the period 1990-1996 but rather come out of the simultaneous fit of the model to data from 1961 through 2001.

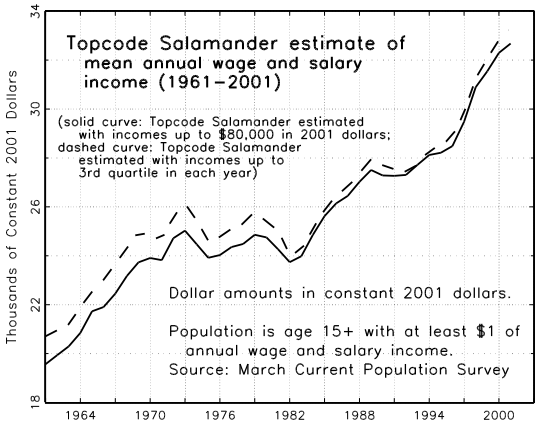

Figure 17

The Topcode Salamander

The Salamander conditioned on education alone achieves moderately close fits very parsimoniously if the distribution it is fitted to is restricted to that of people 25 years of age or older. The whole labor force, as officially defined in the U.S., includes workers ages 16 through 24 as well. Estimating mean annual wage and salary income in the entire labor force requires including these younger workers in the analysis. Their distribution of annual wage and salary income is distinctively different in shape from that of older workers. Their inclusion requires conditioning the Salamander on age as well as education. Further, since the point of including these younger workers is to yield a good estimate of mean annual wage and salary income for comparison with benchmark estimates and the point of this comparison is to establish the adequacy of the Salamander as a model of the far right tail of the distribution truncated by topcoding, the version of the Salamander intended to perform this function is modified to give observations on larger incomes greater weight and to constrain fits to partial distributions and estimates of their means to add to the estimate of the unconditional mean of wage and salary income. This modified version of the Salamander is given the distinctive name of Topcode Salamander to distinguish it from the simpler version of the Salamander introduced as a model of the entire distribution.

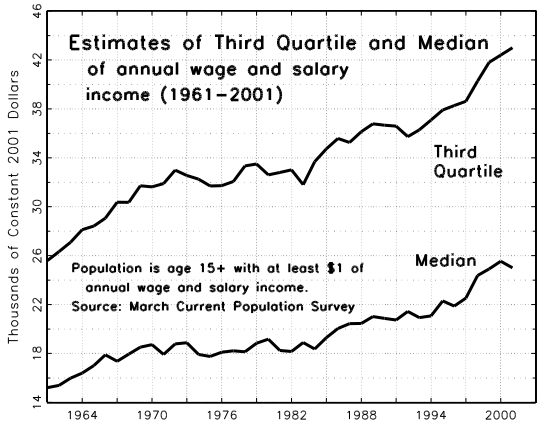

Figure 18

The Topcode Salamander is modified from the basic Salamander as defined by (4) as:

$$g_{it}(x) \equiv \frac{\lambda_{it}^{\alpha_i}}{\Gamma(\alpha_i)} x^{\alpha_i} e^{-\lambda_{it} x} = x \cdot f_{it}(x) \quad (14)$$

where $f_{it}(x)$ is defined by (4). In (14) the subscript i refers to combined age-education groups. There are 12 of these, the product of six education levels and two age groupings. The age groupings are 29 and younger, 30 and older. The integral of (14) from 0 to infinity is (4)'s estimate of mean annual wage and salary income at time t among people in a particular education-age grouping. These sum to the unconditional mean:

$$\overline{x}_t = w_{1t}\ \overline{x}_{1t} + ..... + \ w_{12t}\ \overline{x}_{12t} \quad (15)$$

Each weight is the proportion of the sample total aggregate of wage and salary income received in the form of incomes of $80,000 (in constant 2001 dollars) or less in each of the 12 partial distributions. The Topcode Salamander is fitted to eight relative frequency bins per conditional distribution fitted. These are $1-$10,000, $10,001 to $20,000, up to $70,001 to $80,000. The earlier fit of the regular Salamander was to 18 relative frequencies in bins $5,000 wide from $1 - $5,000 to $85,001 - $90,000. Fewer bins and a smaller maximum income are used with the Topcode Salamander because of the sparseness of the data occasioned by fitting 12 conditional distributions in a year. Because education is not available in the 1963 March CPS' PUMS, the 1962 conditional distributions are interpolated from the 1961 and 1963 conditional distributions. The products, $x_{ijkt}y_{ijkt}$, in each relative frequency bin are fitted. $x_{ijkt}$ is the mean, in year t, of the $k^{th}$ income bin in the conditional distribution corresponding to the $j^{th}$ age group and the $i^{th}$ education level. $y_{ijkt}$ is the corresponding relative frequency.

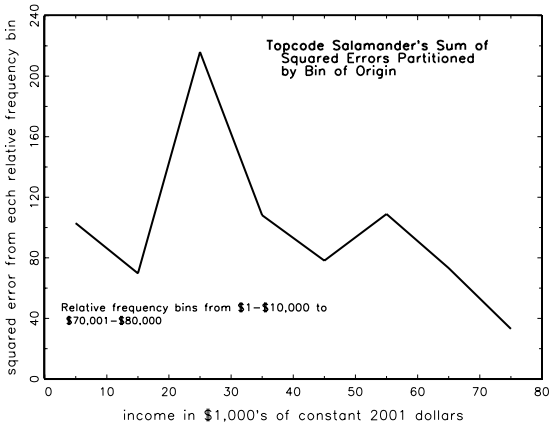

Figure 18

The twelve ω's, one for each education-age distinguished, are given in Table 3. These are estimated by minimizing a weighted sum of squared error via a stochastic search algorithm, describable as simulated annealing. Estimating (14) instead of (4) weights the distribution by income amount, making fit to the central mass of the distribution, where it is best measured (Roemer, 2000: 21), the primary determinant of fit. When (4) is fitted, figure 5 shows that far left tail, where the

distribution is noisy and affected by measurement and other non-sampling errors, is the primary criterion of fit. The analogue of figure 5 for the Topcode Salamander is figure 18, which shows that most of the fit is attributable to the central mass rather than the left tail.

Table 3

| Highest Level of Education | Age | $\omega_i$ Estimated by Fitting Model to All Relative Frequencies | Bootstrapped Standard Error of $\omega_i$ Estimated by Fitting Model to All Relative Frequencies (50 re- samples) |
|---|---|---|---|
| Eighth Grade or Less | 29 or younger | .522551 | .101547 |
| | 30 or older | .377271 | .013341 |
| Some High School | 29 or younger | .547949 | .113489 |
| | 30 or older | .312921 | .007688 |
| High School Graduate | 29 or younger | .491880 | .051049 |
| | 30 or older | .265009 | .004840 |
| Some College | 29 or younger | .570834 | .079653 |
| | 30 or older | .219795 | .003411 |
| College Graduate | 29 or younger | .278555 | .013612 |
| | 30 or older | .173085 | .002387 |
| Post Graduate Education | 29 or younger | .236819 | .009881 |
| | 30 or older | .153477 | .001753 |

The Topcode Salamander's correlation between the fitted and observed dependent variable is .95269 with a standard error of estimate of .003211 estimated via 50 bootstrap replications from the PUMS data through estimation of relative frequencies and the fit of the Topcode Salamander.

**Conclusions**

The March CPS PUMS' minimum topcodeable annual wage and salary income is at a much higher percentile in the nonmetro distribution than in the metro distribution. As you can see in figure 1, the metro distribution has a much heavier (thicker) tail at any given large annual wage and salary income. Setting the minimum topcodeable income to a high percentile makes the PUMS record of the recipient more matchable to a commercial data base of data on the U.S. population. Such a match, if accurate, results in the re-identification of the respondent whose information is contained in the PUMS record. Not only are fewer incomes topcoded when the minimum topcodeable income is at a higher percentile, the protection afforded by topcoding is less. Part of the protection afforded by topcoding is not just the suppression of exact information about a large income, it is the concealment of larger large incomes, tagged as such by their having been topcoded, among

smaller large topcoded incomes, the protection of the herd. Nonmetro PUMS records are also more vulnerable to re-identification than metro ones because the nonmetro population is geographically sparser, e.g., a record identified as one of a person living in a nonmetro county of New Mexico is more readily matched to a commercial database than that of a resident of a metro county of New Mexico. Just a few states account for most of the metro population also making geographic information less identifying of metro cases in a PUMS than nonmetro. Further, the nonmetro population may be less transient and less anonymous locally than the metro population, and hence more easily identified with the help of local knowledge.

The U.S. Bureau of the Census sets a minimum topcodeable annual wage and salary income in terms of nominal dollars often at the unconditional 99$^{th}$ percentile or higher. Currently, the March CPS minimum topcodeable annual wage and salary income is $150,000. As inflation creeps and the distribution changes, this amount has been increased from time to time, often after a decennial census. The minimum topcodeable annual wage and salary income was a nominal $50,000 during the 1970's while rapid inflation lowered it in real terms. By 1980, it had fallen to about the 98$^{th}$ percentile and had biased comparisons of the nonmetro and metro Gini concentration ratios of annual wage and salary income (Angle and Tolbert, 1999). The likely reason that the U.S. Bureau of the Census keeps the minimum topcodeable annual wage and salary income well above the 98$^{th}$ percentile is a concern to prevent topcoding from seriously biasing statistics of annual wage and salary income, even statistics as sensitive to large incomes as the mean or the Gini. As the size and completeness of commercial databases on the U.S. population grow, as data mining and record matching technologies advance, and as the cost of computation decreases, large incomes in the PUMS of Federal surveys become more vulnerable to being matched and re-identified (FCSM, 1994), large nonmetro incomes in particular. A Federal law on the confidentiality of Federal data collection enacted in 2002, the Confidential Information Protection and Statistical Efficiency Act (CIPSEA) requires Federal statistical agencies to re-balance the protection of respondents from disclosure of their identities (the 'CIP' part of the law) with the usefulness of data releases to the public (the 'SE' part of the law). Perhaps the usefulness of the Topcode Salamander in estimating statistics of the whole distribution, as evidenced in the present paper by its estimates of the mean of annual wage and salary income, can ease the Bureau's concern about keeping the minimum topcodeable income so high as a percentile of the distribution. After all, the Salamander yields reasonably good estimates of the mean of annual wage and salary income distribution even when much of the right tail of the distribution has been discarded.

The question of whether the Salamander justifies striking a new balance between protecting respondent confidentiality and maintaining the usefulness of annual wage and salary income data in a PUMS by substantially lowering the minimum topcodeable income turns on the validity of the Salamander as a model of the distribution of annual wage and salary income and how good it is as an estimate of the right tail of the distribution truncated by topcoding. The evidence presented here for the validity and utility of the Salamander are threefold. First, the Salamander fits the conditional distribution of U.S. annual wage and salary income conditioned on education, 1961-2001, with only as many parameters as

there are levels of education distinguished, i.e., with great parsimony. Secondly, a number of visually striking quasi-symmetries in the empirical time series of annual wage and salary income, some undiscussed in the literature, are implied by the algebra of the Salamander.

One of these five pattens in the dynamics of the distribution is however widely discussed in the social science literature. It is the thickening of the distribution over larger annual wage and salary incomes when the mean increases. Since, according to the Salamander, proportional change in a relative frequency depends on its distance from the mean and since the extreme right tail is the farthest point in the distribution from the mean, the Salamander predicts the greatest proportional change in the extreme right tail. So the Salamander implies an apparent suddenness of people new to large wage and salary incomes when mean wage income increases, and a sudden disappearance of many large wage incomes when the mean of annual wage and salary income turns down in a recession. Since mean wage and salary income increases in real terms in most years, there is a continual rapid proportional thickening of the distribution of annual wage and salary income at its rightmost extreme. This thickening has been perceived in the social science literature as a social problem, as the distribution of wage income becoming bimodal, that is, split into a group of rich and a group of poor people. Terms such as 'polarization' or 'a hollowing out of the distribution' have been used to describe this forecast. But the Salamander not only implies a rapid proportional thickening of the wage distribution at its far right end when the mean increases, it implies simultaneously a sizeable proportional decrease in the far left tail of the distribution, the distribution of the smallest wage incomes. In terms of changes in the absolute value of relative frequencies, these decreases in the far left tail are much bigger than those in the far right tail because there are many more very small incomes than very large wage incomes. The relative frequencies in the far left tail move toward the right in the distribution. The Salamander does not imply the emergence of substantial bimodality in the distribution but rather implies that the thickening in the far right tail of the distribution happens simultaneously with beneficial changes in the distribution over all income amounts, the Salamander's version of the familiar saying among economists that "a rising tide", here a rising mean wage, "lifts all boats". As you can see in figure 11 a substantially bimodal wage distribution did not emerge in the U.S., 1961 - 2001. The Salamander provides an explanation for why it did not emerge and why change in the distribution of annual wage and salary income as its mean increases should be welcome.

The third item of evidence for the validity and utility of the Salamander is the comparison of its estimates of the unconditional mean of annual wage and salary income to the estimates that the Census Bureau accepts as benchmarks for the years 1990 through 1996. The Pareto probability density function (pdf) is the standard way of estimating a truncated right tail of a wage distribution, even though its assumption of a constant ratio of truncation point to mean of the truncated right tail is demonstrably false in the U.S. distribution of wage and salary income (cf. Angle and Tolbert, 1999). Nevertheless, besides being traditional, the Pareto pdf is preferred as a model of the right tail because its right tail is heavier (thicker) than that of some pdfs that offer a better fit to the whole distribution such as the lognormal or gamma pdf. It is thought that this

heavier tail is necessary to account for the total aggregate of income received in the truncated right tail, and consequently necessary to estimate mean annual wage and salary income. This concern is behind Quensel's (1944) practice of modeling the left tail and central mass of income distributions with the lognormal pdf while modeling the far right tail with the heavier-tailed Pareto (1897)[9] while arbitrarily choosing the "seam" between the two. The Salamander is a mixture of the gamma pdfs fitted to the distribution of annual wage and salary income conditioned on education. Its mixing weights are the proportions of the labor force at each level of education distinguished. The Salamander's right tail is, generally, heavier (thicker) than that of a gamma pdf fitted directly to the unconditional distribution of annual wage and salary income.

The reason the heaviness of the estimated right tail of the wage distribution is an issue is not concern for finer estimation of the density of large incomes as much as it is a concern with how their truncation affects statistics of the whole distribution sensitive to the far right tail of the distribution. Perhaps the most important of such statistics is mean annual wage and salary income. There is a benchmark against which to compare Salamander estimates of this statistic. The Salamander estimates of mean annual wage and salary income are better in the years 1990-1996 than the Census Bureau's estimates based on untopcoded March CPS data (Roemer, 2000), even though these Salamander estimates come out of the fit of the Salamander to March CPS annual wage and salary income data from 1961 through 2001. The Salamander thus passes the test of whether its right tail is sufficiently heavy to produce a correct estimate of the mean.

Summarizing the three types of evidence presented in this paper about the Salamander, one can say that its estimates of the annual wage and salary income distribution are more than just close but mimic the distribution's dynamics. The Salamander does so parsimoniously and robustly over many decades. The Salamander can be well estimated without observations in the far right tail of the distribution, freeing Federal Statistical agencies to lower their minimum topcodeable annual wage and salary income. Even estimated with incomes at or below the 3rd quartile (75th percentile), the Salamander provides excellent estimates of mean annual wage and salary income according to benchmark standards of the Census Bureau (Roemer, 2000). The Salamander is a candidate for closer examination as a "symbolic statistical object" (Billard and Diday, 2003), a generator of statistics on annual wage and salary income distribution. This paper should be read as distinctly raising the hope that the Salamander may be a substitute for sample observations on very large annual wage and salary incomes. If the increasing risk of the disclosure of the identities of respondents reporting large annual wage and salary incomes in Federal household surveys prompts restriction of the PUMS of these surveys via a lowering of the minimum topcodeable income, the Salamander may allow users of these PUMS to continue to estimate statistics of the distribution.

---

[9] After making a claim that the lognormal pdf is the natural functional form of income distributions (Montroll and Badger, 1974 ), Badger (1980) modified this claim to a functional form that approximates the lognormal in its left tail and central mass and a Pareto pdf in its right tail in recongition of the practice of economists of only using the lognormal only to model the left tail and cental mass of income distributions while using the Pareto pdf, with its heavier right tail, to model the right tail of income distributions. Badger cites Quensel (1944 ) as beginning that practice. The Pareto pdf has been the standard pdf model of the truncated right tail of income distributions at the U.S. Bureau of the Census (Miller, 1966; Henson, 1967; Shryock and Siegel, 1973; and Spiers, 1977). It is also has been traditionally favored by sociologists for the same application (Parker and Fenwick, 1983).

**APPENDIX A: Data and Methods**

The distribution of wage and salary income is estimated with data from the March Current Population Surveys (1962-2002). The March Current Population Survey (CPS) is known as the Annual Demographic Survey. It has a supplementary questionnaire which includes questions on income received in the previous calendar year, posed on behalf of the U.S. Bureau of Labor Statistics. The CPS is conducted monthly by the U.S. Bureau of the Census (Weinberg, Nelson, Roemer, and Welniak, 1999). The CPS has a substantial number of households in its nationwide sample. The Salamander is first estimated in the present paper on the population 25 + in age, earning at least $1 in annual wage and salary income, and then for comparability with Roemer's (2000) implied estimates of mean annual wage and salary income, 1990-1996, the version of the Salamander called the Topcode Salamander is estimated on the population 16+ in age, earning at least $1 in annual wage and salary income. The age restriction to 25+ is to allow the more educated to be comparable to the less educated. When the population of wage earners ages 16 to 24 is included in the recipient population, age has to be explicitly incorporated into the Salamander along with education with a resulting loss of parsimony. Both versions of the Salamander are fitted by a stochastic search algorithm, a variety of simulated annealing (Kirkpatrick, Gelatt,and Vecchi. 1983). The data of the March CPS of 1962 through 2002 used in the present paper were purchased from Unicon, inc. (Unicon, inc, 2002; Current Population Surveys, March 1962-2002), which provides the services of data cleaning and extraction software, along with substantial research on variable definitions and comparability over time. Unicon, inc was not able to find a copy of the March 1963 CPS which contains data on education. Consequently, the distribution of wage and salary income received in 1962 (from the March 1963 CPS) conditioned on education is interpolated from the 1961 and 1963 (from the 1962 and 1964 March CPS').

All dollar amounts in the March CPS' are converted to constant 2001 dollars using the PCE (personal consumption expenditure) price index numbers form Table B-7 Chain-type price indexes for gross domestic product, **Economic Report to the President**, February 2003 (Council of Economic Advisers, 2003).

The numbers of persons in the March Current Population Survey in each year and the number of them meeting the criterion for selection are:

| March CPS of | Total number of person records in the March Current Population Survey | people, age 25+, who earned at least $1 in previous calendar year |
|---|---|---|
| 1962 | 71,745 | 22,923 |
| 1963 | 54,282 | 15,147 |
| 1964 | 54,543 | 23,903 |
| 1965 | 54,516 | 23,839 |
| 1966 | 110,055 | 46,656 |
| 1967 | 104,902 | 45,266 |
| 1968 | 150,913 | 47,157 |
| 1969 | 151,848 | 48,088 |
| 1970 | 145,023 | 46,004 |
| 1971 | 147,189 | 46,088 |
| 1972 | 140,432 | 44,143 |
| 1973 | 136,221 | 43,200 |
| 1974 | 133,282 | 43,043 |
| 1975 | 130,124 | 42,424 |
| 1976 | 135,351 | 43,888 |
| 1977 | 160,799 | 52,663 |
| 1978 | 155,706 | 52,255 |
| 1979 | 154,593 | 52,793 |
| 1980 | 181,488 | 63,429 |
| 1981 | 181,358 | 64,108 |
| 1982 | 162,703 | 57,877 |
| 1983 | 162,635 | 57,995 |
| 1984 | 161,167 | 58,049 |
| 1985 | 161,362 | 59,819 |
| 1986 | 157,661 | 59,596 |
| 1987 | 155,468 | 59,603 |
| 1988 | 155,906 | 60,501 |

| | | |
|---|---|---|
| 1989 | 144,687 | 57,158 |
| 1990 | 158,079 | 62,883 |
| 1991 | 158,477 | 62,942 |
| 1992 | 155,796 | 62,085 |
| 1993 | 155,197 | 61,331 |
| 1994 | 150,943 | 59,575 |
| 1995 | 149,642 | 59,999 |
| 1996 | 130,476 | 53,358 |
| 1997 | 131,854 | 54,553 |
| 1998 | 131,617 | 54,056 |
| 1999 | 132,324 | 54,659 |
| 2000 | 133,710 | 55,925 |
| 2001 | 128,821 | 53,967 |
| 2002 | 217,219 | 89,200 |

The measurement of education changed in the CPS after the 1990 Census from a count of years of school completed to a more degree oriented measure which better measures the diversity of post-secondary education. The present study reconciles the two categorizations of educational attainment by collapsing both sets of categories to an ordinal polytomy of five categories. The crudeness of this categorization obliterates the distinction between the two different categorizations of educational attainment. The categories of highest level of education attained used here are:

| |
|---|
| elementary school or less |
| some high school |
| completed four years of high school |
| some college |
| completed four or more years of post-secondary education |

This paper estimates the distribution of annual nonmetro wage and salary income the traditional way, in terms of relative frequencies of observations falling into bins of fixed width. There are many ways to estimate a distribution. All of them involve a trade-off between parsimony of model and error of fit. Parsimony is expressed in the amount of smoothing of

the estimate. In terms of fixed bins, the greater the bin width, the fewer bins are used, and the greater the degree of aggregation and the smoother the estimate of the distribution. A wage and salary income distribution of a large population defined in geographic terms has been a familiar statistical object for over a century. It is known to be right skewed (Pareto's Law, broadly construed) and usually unimodal after smoothing. Angle (1994) demonstrates the existence of a micro-structure of frequency spikes over round income amounts in March CPS income data, indicative of respondents reporting incomes to Census Bureau interviewers with fewer significant digits than asked for. Census Bureau questionnaires ask for incomes to the nearest $1. Angle (1994) shows that this rounding of income amounts does not introduce a net upward or downward bias. In published tabulations, the Census Bureau, traditionally, presents income distributions near their mode in terms of relative frequencies in bins of fixed length, e.g., $5,000, and in the increasingly sparse right tail, in terms of bins of increasing width. This policy is intended to keep the standard errors of estimate of the right tail bins comparable to those of the bins near the mode. However, such presentation disguises how right skewed income distributions are because it is difficult to mentally adjust the relative frequency for the increasing bin length in the right tail. The present paper estimates a distribution with relative frequency bins that are fixed length, either $5,000 or $10,000 wide (in terms of constant 2001 dollars), to facilitate comparison between the more dense left tail and the less dense right tail.

**Appendix B: Derivation of Partial Derivative of $f_{it}(x)$ with respect to $\bar{x}_t$, $\partial f_{it}(x) / \partial \bar{x}_t$**

The partial derivative of $f_{it}(x)$ with respect to $\bar{x}_t$ gives an expression for how $f_{it}(x)$ changes as a function of $\bar{x}_t$.

$$f_{it}(x) = \frac{\lambda_{it}^{\alpha_i}}{\Gamma(\alpha_i)} x^{\alpha_i - 1} e^{-\lambda_{it} x}$$

$$= \exp\left[\alpha_i \ln(\lambda_{it}) - \ln(\Gamma(\alpha_i)) + (\alpha_i - 1)\ln(x) - \lambda_{it} x\right]$$

where,

$$\lambda_t \approx \frac{(1-\omega_i) \sum_{i=1}^{I} \frac{w_{it}}{\omega_i}}{\bar{x}_t}$$

and

$$\alpha_i \approx \frac{(1-\omega_i)}{\omega_i}$$

so:

$$\frac{\partial f_{it}(x)}{\partial \bar{x}_t} = \exp\left[\alpha_i \ln(\lambda_{it}) - \ln(\Gamma(\alpha_i)) + (\alpha_i - 1)\ln(x) - \lambda_{it} x\right] \cdot$$

$$\frac{\partial}{\partial \bar{x}_t}\left[\alpha_i \ln(\lambda_{it}) - \ln(\Gamma(\alpha_i)) + (\alpha_i - 1)\ln(x) - \lambda_{it} x\right]$$

$$= f_{it}(x) \cdot \frac{\partial}{\partial \bar{x}_t}\left[\alpha_i \ln(\lambda_{it}) - \ln(\Gamma(\alpha_i)) + (\alpha_i - 1)\ln(x) - \lambda_{it} x\right]$$

and since the non-zero terms in the argument of the partial derivative are:

$$\frac{\partial\left(\alpha_i \ln(\lambda_{it})\right)}{\partial \bar{x}_t} = \alpha_i \frac{\partial}{\partial \bar{x}_t}\left[\ln(1-\omega_i) + \ln\left(\sum_{i=1}^{I} \frac{w_{it}}{\omega_i}\right) - \ln(\bar{x}_t)\right]$$

$$= -\alpha_i \frac{\partial}{\partial \bar{x}_t}\left[\ln(\bar{x}_t)\right]$$

$$= -\frac{\alpha_i}{\bar{x}_t}$$

and:

$$\frac{\partial\left(\lambda_{it}\, x\right)}{\partial \bar{x}_t} = (1-\omega_i)\left(\sum_{i=1}^{I} \frac{w_{it}}{\omega_i}\right) x \frac{\partial}{\partial \bar{x}_t}\left(\frac{1}{\bar{x}_t}\right)$$

$$= -(1-\omega_i)\left(\sum_{i=1}^{I} \frac{w_{it}}{\omega_i}\right) x \left(\frac{1}{\bar{x}_t^2}\right)$$

$$= \frac{-\lambda_{it}\, x}{\bar{x}_t}$$

it follows that:

$$\frac{\partial f_{it}(x)}{\partial \bar{x}_t} = f_{it}(x) \cdot \left[ -\frac{\alpha_i}{\bar{x}_t} - \left( -\frac{\lambda_{it}\, x}{\bar{x}_t} \right) \right]$$

$$= f_{it}(x) \cdot \left[ \frac{-\alpha_i + \lambda_{it}\, x}{\bar{x}_t} \right]$$

and because in the model:

$$\bar{x}_{it} = \frac{\alpha_i}{\lambda_{it}}$$

the partial derivative of $f_{it}(x)$ with respect to $\bar{x}_t$ is:

$$\frac{\partial f_{it}(x)}{\partial \bar{x}_t} = f_{it}(x) \cdot \lambda_{it} \cdot \left[ \frac{x - \bar{x}_{it}}{\bar{x}_t} \right] \qquad (10)$$

**Appendix C: The Salamander's Explanation for Why Standardizing Relative Frequencies Equidistant from the the Mean of Wage and Salary Income and Reversing the Sign of One Makes the Graphs of the Two Transformed Relative Frequencies Overlap**

The Salamander is a static model unless driven by an exogenous source of change. Two such sources are modeled, one, the unconditional mean changes proportionally much more quickly than the other, and because of the algebra of the Salamander also has a larger effect. Thus it is possible to approximate change in the relative frequency of the Salamander at income x by applying Newton's approximation taking change in the unconditional mean of wage and salary income alone into account:

$$f_{i1}(x) \approx f_{i0}(x) + \frac{\partial f_{i0}(x)}{\partial \bar{x}_0} \Delta \bar{x}_1$$

$$f_{i2}(x) \approx f_{i1}(x) + \frac{\partial f_{i1}(x)}{\partial \bar{x}_1} \Delta \bar{x}_2$$

$$\vdots \quad \vdots$$

$$\vdots \quad \vdots$$

which implies, given (10):

$$f_{i1}(x) \approx f_{i0}(x) + f_{i0}(x)\, \lambda_{i0} \left( \frac{x - \bar{x}_{i0}}{\bar{x}_0} \right) \Delta \bar{x}_1$$

$$f_{i2}(x) \approx f_{i0}(x) \left( 1 + \lambda_{i0} \left( \frac{x - \bar{x}_{i0}}{\bar{x}_0} \right) \Delta \bar{x}_1 \right) \left( 1 + \lambda_{i1} \left( \frac{x - \bar{x}_{i1}}{\bar{x}_1} \right) \Delta \bar{x}_2 \right)$$

$$\vdots \quad \vdots$$

$$\vdots \quad \vdots$$

When the $f_{it}(x)$'s are standardized, $f_{i0}(x)$ cancels out of the numerator and denominator of all of them, leaving an expression for the standardized $f_{it}(x)$'s that is identical for incomes, $x_{il}$ and $x_{ir}$, equidistant to the left and right of the conditional mean income, $\bar{x}_{it}$, except for the sign of each $(x - \bar{x}_{it})$ term. Thus by reversing the sign of one set of $f_{it}(x)$'s, either the set in the left tail or that of the right, the two sets of terms are made identical, under the assumption that the $\Delta \bar{x}_t$'s are the sole source of change. This finding for the conditional distributions aggregate up to the unconditional distribution via the same weighted summation as the summation of the conditional means to the unconditional mean.

**References**

Angle, John. 1993. "An apparent invariance of the size distribution of personal income conditioned on education." **1993 Proceedings of the Social Statistics Section of the American Statistical Association**. Pp. 197-202.

_____. 1994. "Frequency spikes in income distributions." **1994 Proceedings of the Business and Economic Statistics Section of the American Statistical Association**. 265-270.

_____. 2002a. "Modeling the dynamics of the nonmetro distribution of wage and salary income as a function of its mean". **2002 Proceedings of the American Statistical Association, Business and Economic Statistics Section. Alexandria, VA: American Statistical Association.**

_____. 2002b. "The statistical signature of pervasive competition on wage and salary incomes." **Journal of Mathematical Sociology** 26: 217-270.

_____. forthcoming. "The dynamics of the nonmetro distribution of wage and salary income". **Estadistica**.

_____ and Charles Tolbert. 1999. "Topcodes and the Great U-Turn in Nonmetro/Metro Wage and Salary Income Inequality". **Staff Report** (#9904). Washington, DC: Economic Research Service, U.S. Department of Agriculture.

Badger, W. 1980. "An entropy-utility model for the size distribution of Income". Pp. 87-120 in B.J. West, (ed.), **Mathematical Models as a Tool for the Social Sciences**. New York: Gordon and Breach.

Bauer, E. and R. Kohavi. 1999. "An empirical comparison of voting classification algorithms: bagging, boosting, and variants". **Machine Learning** pp. 105-142.

Billard, L. and E. Diday. 2003. "From the statistics of data to the statistics of knowledge: symbolic data analysis". **Journal of the American Statistical Association** 98: 470-487.

Boisvert, R. 1977. "An analysis of income distribution in the northeast using the gamma density". **Cornell Agricultural Economics Staff Paper** #77-20. Ithaca, NY: Cornell University.

Cowell, F. 1977. **Measuring Inequality: Techniques for the Social Sciences**. New York: Wiley.

Current Population Surveys, March 1962-2002. [machine readable data files]/ conducted by the Bureau of the Census for the Bureau of Labor Statistics. Washington, DC: U.S. Bureau of the Census [producer and distributor], 1962-2002. Santa Monica, CA: Unicon Research Corporation [producer and distributor of CPS Utilities], 2002.

Council of Economic Advisers. 2002. **Economic Report to the President**. Washington, DC: U.S. Government Printing Office.

Dagum, C. 1977. "A new model of personal income distribution: specification and estimation". **Economie Appliquée** 30:413-437.

Economic Research Service. 2004. "Measuring Rurality" (http://www.ers.usda.gov/briefing/rurality).

FCSM, Subcommittee on Disclosure Limitation Methodology, Federal Committee on Statistical Methodology. 1994. **Report on Statistical Disclosure Limitation Methodology, (Statistical Policy Working Paper #22)**. Washington, DC: U.S. Office of Management and Budget, Statistical Policy Office, Federal Committee on Statistical Methodology,

Henson, Mary. 1967. **Trends in the Income of Families and Persons in the United States: 1947 to 1960**. Technical Paper # 17. Washington, DC: U.S. Bureau of the Census.

Kirkpatrick, S., C.D. Gelatt,and M.P. Vecchi. 1983. "Optimization by simulated annealing." **Science** 220: 671-680.

March, L. 1898. "Quelques exemples de distribution de salaires". **Journal de la Société Statistique de Paris**. June and July. Pp. 193-206, 241-248 (found via Dagum., 1977).

McDonald, J. and B. Jensen. 1979. "An analysis of some properties of alternative measures of income inequality based on the gamma distribution function." **Journal of the American Statistical Association** 74: 856-860.

Miller, Herman. 1966. **Income Distribution in the United States**. Washington, DC: U.S. Bureau of the Census.

Montroll, E.W. and W. Badger. 1974. **Introduction to Quantitative Aspects of Social Phenomena**. New York: Gordon and Breach.

Morris, Martina, Annette Bernhardt, and Mark Handcock. 1994. "Economic Inequality: New Methods for New Trends". **American Sociological Review**
59: 205-219.

Pareto, Vilfredo. 1897. **Cours d'Économie Politique**. Vol.2. Lausanne: Rouge

Parker, Robert and Rudy Fenwick. 1983. "The Pareto Curve and its utility for open-ended income distributions in survey research". **Social Forces** 61:872-885.

Peterson, L. and H. von Foerster. 1971. "Cybernetics of taxation: the optimization of economic participation". **Journal of Cybernetics** 1:5-22.

Pollack, Andrew. "Missing limb? Salamander may have the answer." **New York Times** (Late New York Edition), Tuesday, September 24, 2002, pp F1, F4.

Quensel, C.E. 1944. **Inkomstfoerdelning och skattetryck**. Stockholm: Sveriges Industriforbund (Cited in Badger, 1980.).

Roemer, Marc. 2000. "Assessing the Quality of the March Current Population Survey and the Survey of Income and Program Participation Income Estimates, 1990-1996". Washington, DC: U.S. Census Bureau. http://www.census.gov/hhes/www/income/assess1.pdf.

Salem, A. and T. Mount. 1974. "A convenient descriptive model of income distribution: the gamma density". **Econometrica** 42: 1115-1127.

Shorrocks, A. 1975. "On stochastic models of size distributions". **The Review of Economic Studies** 42:631-641.

Shryock, Henry, Jacob Siegel, and associates. 1973. **The Methods and Materials of Demography**. Vol.1. Washington, DC: U.S. Bureau of the Census.

Spiers, Emmett. 1977. "Estimation of summary measures of income size distribution from grouped data". **Proceedings of the Social Statistics Section of the American Statistical Association.** Pp. 252-277.

Unicon, inc. 2002. **Manual for March Current Population Surveys.** Santa Monica, CA: Unicon

U.S. Bureau of Economic Analysis. 2003. Table # CA34 - wage and salary summary estimates. Internet address: 'http://www.bea.doc.gov/bea/regional/reis/default.cfm#$2' .

Weinberg, Daniel, Charles Nelson, Marc Roemer, and Edward Welniak. 1999. "Fifty years of U.S. income data from the Current Population Survey". **American Economic Review** vol.89, issue 2 (Papers and Proceedings of the 111th Annual Meeting of the American Economic Association), 18-22.