Statistical Literacy and the Lognormal Distribution

Milo Schield

Augsburg University; Minneapolis, Minnesota

Abstract

Income inequality is a hot topic. There is no public data on the income share of the top 1% of households; those percentages are estimates. Those estimates vary from 5% to 40%. They are based on different data using different definitions and different models. This paper studies a common model: the log-normal distribution. This paper shows six things for any log-normal distribution. (1) The mean-median ratio determines the shape of the distribution, the income share of the top 1%, and the Gini coefficient. (2) The log-normal model is a fairly good fit to the actual distribution of 2016 US household incomes. (3) If the distribution of subjects by income is log-normal, then the distribution of total income by household income will also be log-normal. (4) The percentage of subjects that have below-average incomes always equals the percentage of total income that is earned by those with above-average income. (5) These below/above percentages can be used to measure income inequality in a way that is more accessible than the Gini coefficient. (6) The product of a normal and a log-normal can be modeled as the product of two log-normal distributions. Including the log-normal distribution in a statistical literacy class helps decision makers focus on essentials.

Key Words: Gini coefficient, Relative share ratio, Income inequality, Top 1%, statistical education

1. Income Inequality

Income inequality is a hot topic politically. The most important numbers are the income shares of the super-rich. But these numbers are not publicly available. The publicly touted high-income shares are estimates: estimates based on different data sources, different definitions of income and using different models.

Using different definitions of income has a big influence on the income share of the top 1%. Historically, the definition of income was regular income: income excluding capital gains. Using this definition, Piketty (2013) had estimated the income share of the top 1% of US households was 20%. After changing the definition to include capital gains, Piketty et al; (2018) estimated the income share of the top 1% was 40%.

Another influence is the choice of model: lognormal versus Pareto. Piketty modeled using the Pareto. After providing specific estimates of the income share of the top 1% in various countries, he argued that increasing income inequality was bad and required governmental intervention. Piketty's book, *Capital in the Twenty-First Century*, has sold over 2.5 million copies. Paul Krugman (2014) christened *Capital* as the "most important book of the year – and maybe of the decade." For details, see Appendix A.

Sommeiller and Price (2018) went even further in generating statistics showing that income inequality is substantial – especially among the top 1%, the top 0.1% and the top 0.001%. Since this high-end data is top-coded and not available, they followed Piketty and obtained it from a model of the high-end of the income distribution: the Pareto distribution. For details on the Pareto distribution see Appendix B

Since these top income shares involved very specific numbers, readers typically assumed they were summaries from actual data. They are not. Income data is typically top-coded. These income shares were the result of modeling: modelling using a distribution. See Appendix A for more details.

Distributions typically describe, but they can be used to predict. Distributions are worth understanding because they simplify the complexity of real-world data. They allow decision-makers to focus on essentials. This paper focuses on a particular model of income: the lognormal distribution.

2. *Lognormal Distribution of Subjects*

Aitchison and Brown (1957, page 1): "In many ways, it [the Lognormal] has remained the Cinderella of distributions, the interest of writers in the learned journals being curiously sporadic and that of the authors of statistical test-books but faintly aroused." "We ... state our belief that the lognormal is as fundamental a distribution in statistics as is the normal, despite the stigma of the derivative nature of its name."

For background on the lognormal, see Limpert et al (2001). Aitchison and Brown (1957) reviewed the nature and general properties of the lognormal. Let μ and σ signify the mean and standard deviation of the underlying normal. Let Mean, Median and StdDev signify the properties of the resulting lognormal. Mathematicians usually focus on the underlying normal: μ and σ. Decision-makers are more comfortable starting with the summary statistics: Mean, Median and StdDev. Appendix C shows how the underlying parameters can be derived from the lognormal summary statistics.

Solving for μ and σ in terms of the mean and median gives:
Eq. 1 $\mu = Ln(median)$
Eq. 2 $\sigma^2 = 2*Ln(mean/median)$

Solving for μ and σ in terms of the mean and standard deviation gives:
Eq. 3 $\mu = Ln(median)$
Eq. 4 $\sigma = SQRT\{Ln[(StdDev/Mean)\char`^2 \ + \ 1]\}$

It will be shown that the distribution of income shares is entirely determined by the shape of a lognormal – which is entirely determined by σ. As noted above, σ is entirely determined by either of two ratios: the mean to median ratio or the standard deviation to mean ratio (the coefficient of variation). These ratios are most convenient; they eliminate the need to adjust for inflation or exchange rates.

3. *Lognormal Distribution of Households: Fit to Observed Distribution*

The lognormal distribution is commonly used to model the distribution of households by income. This fit is shown for the distribution of households generated by the US Census Current Population Survey (2017).
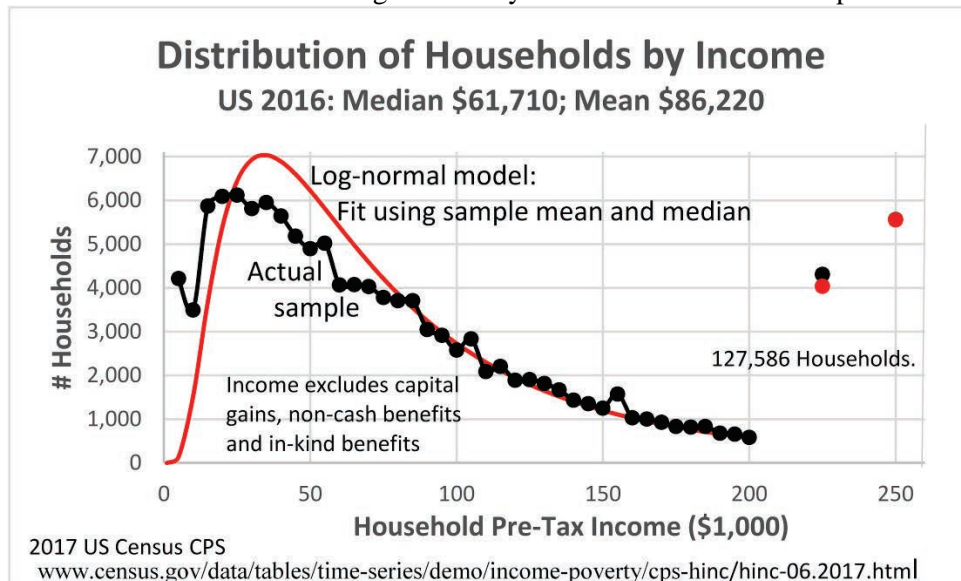

Figure 1: US Distribution of Income: Actual vs. Lognormal model

The lognormal fits well for household incomes above the mean: $86,220. Below there are two problems. First the higher number of households (relative to the lognormal model) with incomes of less than $10,000. Second the lower number of households between the mode and the mean. Both may be related to the CPS standard definition of income.

CPS income is pretax and includes social security, workers compensation, retirement income and net business income. Income excludes lump-sum receipts, capital gains and sales of assets. Income excludes non-cash or in-kind benefits (e.g., employer-provided health insurance, Medicare, Medicaid, and food stamps). Finally, CPS income is gross – not net; it excludes expenses for social security and Medicaid.

Omitting non-cash and in-kind benefits may explain the high number of households with "income" of less than $10,000. If these income alternatives were included, that might push these very-low income households into incomes between the mode and the median.

The two dots at the $225K income represent the number of households having incomes between $200K and $250k. Notice the small separation between the sampled actual (black dot) and the lognormally modelled (red dot). The single red dot at $250K represents the total number of households having incomes greater than $250k. The black dot (representing the total number of households greater than $250k) is completely hidden underneath the associated red dot (the modelled number having income over $250k).

Among US households, 7.5% have incomes over $200k (4.3% have incomes over $250k). Details on these high-income households are almost invisible in this CPS public data due to top-coding for privacy.

The CPS also has an alternate definition of income that includes income sources commonly available to those with below average incomes. It may be that the lognormal model will do a better job of fitting that more-inclusive definition of income.

4. History of the CPS Mean Median Ratio in the US

Table 1 gives this history of the US Census CPS mean-median ratio in pre-tax, current dollars. The data for 1980 to 1995 is from the US Statistical Abstract (1998), Table 741. The newer data is from the US Census CPS Supplement surveys. The continuing increase in the mean-median ratio is obvious.

Mean	Median*	Ratio	Year	Mean	Median*	Ratio	Year
86,220	61,713	1.40	2017	44,938	34,076	1.32	1995
75,738	53,984	1.40	2014	37,403	29,943	1.25	1990
67,392	49,355	1.37	2010	29,066	23,618	1.23	1985
63,344	47,650	1.33	2005	21,063	17,710	1.19	1980

Table 1: Mean & Median US CPS Household Incomes (current dollars)

* Medians obtained by linear interpolation from grouped data.

5. Statistical Literacy: Mean-Median Ratio

Statistical literacy requires an analysis of the causes and effects of the levels and changes in this ratio. What level of inequality is desirable? Few have addressed this issue. It is easier to comment on changes. Typically, an increase in inequality is presumed to be bad; a decrease is presumed to be good. See Wilkinson and Pickett (2009). Certainly income involving illegal activities, crony capitalism and special political favors is generally bad. Few attempt to argue that inequality is good. See Dorfman (2017).

A balanced presentation should include a discussion of those virtues that generate income inequality. Good examples include new technology, increased access to venture capital, assortative mating, mass media, the internet, lotteries and globalization. A balanced presentation should also include those factors that prevent the poor from using their talents to advance. Depending on the country, these might include the inability to own land, the lack of a fair system of justice to enforce property rights and contracts, crime and warring factions, unstable currency and inability to start a small business without political approval/connections.

A balanced presentation should review the consequences – both costs and benefits – of income inequality.
- Costs: these may include envy, lack of social cohesion and social unrest and chaos. They may include poorer health, more crime and shorter lifespan. See Wilkenson and Pickett (2009)

- Benefits: In California 48% of the state taxes are paid by those in the top 1%. (Ashton, 2018). For the 2014 US Personal Income tax, "The top 1 percent paid a greater share of individual income taxes (39.5%) than the bottom 90 percent combined (29.1%)." (Greenberg, 2016). The top 20 US philanthropists donated over 10 billion in 2016. (Wang, 2017). Income inequality offers greater incentives for the poor to work harder (e.g., to start a business) to escape poverty.

A balanced presentation might also consider the income distribution after redistribution: after government-related fees (e.g., taxes, social security and Medicare deductions); after government-related income and assistance (e.g., Social security disability, Medicare, worker's compensation, welfare, food stamps). See Sammartino, F. (2017) and Scheidel (2017).

But rather than starting with external (exogenous) influences, costs and benefits, start with those factors for which there is income data.

6. What Explains the Distribution of Household Income in the US

Table 2 and Table 3 show CPS mean incomes for different groups. The bigger the difference in incomes within a group or factor, the stronger the influence that factor may have on income inequality.

NUMBER OF EARNERS

Characteristic	Median income	Mean income	% Col Number
No earners	23,000	**35,000**	24%
One earner	51,000	71,000	37%
Two or more	104,000	132,000	40%
Two earners	98,000	**125,000**	31%
Three earners	122,000	147,000	6%
Four or more	151,000	171,000	2%

Table 2: Household Income by # of Earners

Differences in 2017 US Mean Incomes By

Dollars	Household Characteristic
$90,000	Education of Householder [BA vs HS]
90,000	Number of Earners [Two vs none]
77,000	Type of Household [Marital; Gender]
75,000	Size of Household
62,000	Work Status of Householder [FT/PT]
49,000	Age of Householder
41,000	Tenure [Housing: Own vs. Rent]
28,000	Type of Residence [City vs Rural]
23,000	Region/Divisions [of US]

Table 3: Income Differences by Characteristic

The difference in mean incomes within a characteristic is a crude way of estimating the correlation of that characteristic with income. The bigger the income difference, the higher the raw (total) correlation.

Table 2 shows the incomes associated with the various numbers of earners. Table 3 shows the income differences by characteristic. To avoid large differences involving a small fraction of the population, groups are chosen that have about 5% prevalence or better. Thus, the income difference for number of earners compares two earner families with no earner families. The income difference for education compares at-least Bachelors with just high school.

Education and number of earners are the two main influences on income inequality within these factors.

Several of these household characteristics may be highly correlated. Education may be correlated with the number of earners, the type and size of household, the work status of the householder, the housing tenure and the residence (city vs. rural). Thus, the income differences associated with the top two characteristics (education and number of earners) may explain much – if not most – of the variation in household incomes.

This data was obtained from the Social and Economic Supplement to the US Census Current Population Survey (2017). Schield (2018). Obtaining partial correlations requires access to underlying micro-data

7. Lognormal Distribution of Incomes

The distribution of subjects by income is generally modelled by the log-normal. The log-normal distribution arises naturally from any random process where the randomness is multiplicative (just as the Normal distribution arises naturally from any random process where the randomness is additive).

If the subjects are distributed lognormally by X, we may not care about the distribution of X. This is certainly true in modelling weight, the length of texts or the size of cities. But if X is monetary (income, wealth or financial losses), many do care. E.g., what fraction of the total income is earned by the top 1%?

Aitchison and Brown (1957, page 12 theorem 2.6) presented two little-known properties of such distributions. They showed that if f(x) is lognormal with underlying Normal parameters μ and σ then x*f(x) is also lognormal with Normal parameters of μ1 and σ1 where

Eq. 5 $μ1 = μ + σ^2$ $σ1 = σ$

This extremely important relationship is currently not included in the log-normal entries for Wikipedia, Wolfram's MathWorld or MathWorks. A three-year web search gave only one match: Irfan (2014).

Figure 2 illustrates the probability distribution functions (PDFs) for both distributions. Both PDFs are rescaled for visibility. Both are shown as a percentage of the maximum for the household distribution. For any pair of subject and income lognormals, their PDFs intersect at the average income per household.
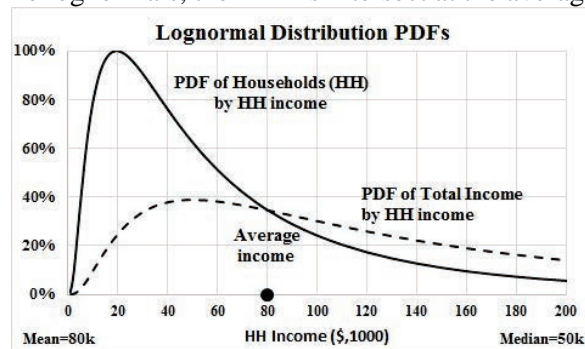

Figure 2: Distribution of HHs and Total Income by HH Income

CDF$ is used to describe the cumulative distribution of total income by household (HH) income. The mode of the income distribution is always the same as the median of the subject distribution ($50k in this case); the median and mean of the income distribution are $128k and $205k. Half of all the income is related to households with less than $128K of income, 75% below $246K and 95% below $608k. Less than 2% of total income is related to households with over one million in income. As a percentage of total income, these high-income households don't have that much income. See Appendix I for details.

In this lognormal model (mean = $50k; median = $80k), the top 1% of households account for 8.9% of total income. According to Wikipedia, the top 1% received 14.6% of the CBO pre-tax income in 2011; up from 13.3% in 2009. This four point difference could indicate that the CPS mean is too low, that the US pre-tax incomes are not lognormal or that different definitions and/or data are being used.

8. *Income Distribution and Excess Income*

Appendix D presents the results of a lognormal model using the mean and median values for the 2017 Census CPS data. The top 1% of households (above $414K) have 6.6% of the total income: almost seven times their equal share. The top 0.1% (above $772K) have 1.2% of the total income: 12 times their equal share. The top 0.01% (above $1,292K) have 0.19% of the total income: 19 times their equal share. One problem with taxing the very rich (on income), is that there are so few of them.

If these income share percentages seem low, perhaps income and wealth are being confused. E.g., the top 1% of households controls 40% of US wealth. DQYDJ (2018) provides comparable results using the actual data for both income and wealth.

9. Lognormal Income Balance

When subjects are distributed lognormally by income, Aitchison and Brown (1957, p. 113) showed that "the proportion of persons with less than the mean income is the complement of the proportion of income held by these persons."

Eq. 6   CDF(mean) = 1 – CDF$(mean)

If subjects are distributed lognormally by income and 75% have below-average incomes, it follows that this group has 25% of the total income. Alternatively, 25% of the subjects have above-average income and this group has 75% of the total income.

This Aitchison-Brown (A-B) income balance between the share of subjects and the share of income for any lognormal distribution is simple, important and deserves to be better known. In Figure 2, the balance condition holds that the fraction of the subject distribution (solid line) that is below the average income always equals the fraction of the income distribution (dashed line) that is above the average income. This equality is plausible, but not obvious.

Appendix E presents the graphical argument given by Aitchison and Brown (1957) using the Lorenz curve. Appendix F presents an algebraic argument similar to the one given by Aitchison and Brown (1957). Appendix G presents a graphical proof by the author using the underlying normal distributions.

The percentage of subjects that have below-average incomes is a simple way of measuring income inequality. For a lognormal distribution, this percentage is identical to the percentage of total income that is associated with subjects having above-average incomes. The higher these percentages, the greater the income inequality. They may be overly simplistic, but they are readily understandable: more than a Gini.

The income share of those with above average incomes is always equal to the fraction of households with below average incomes. These two fractions are entirely determined by the mean-median ratio. The formula for that summary statistic is given by:
Eq. 7   CDF(Mean) = [Erf{Sqrt[Ln(Mean/Median) /2] } + 1] / 2
Eq. 8   CDF(Mean) = Phi{Sqrt[Ln(Mean/Median) /2] }

Figure 3 shows the percentage of subjects that have below-average incomes as a function of the mean-median ratio. As noted, this percentage equals the percentage of total income that is associated with households having above-average incomes.
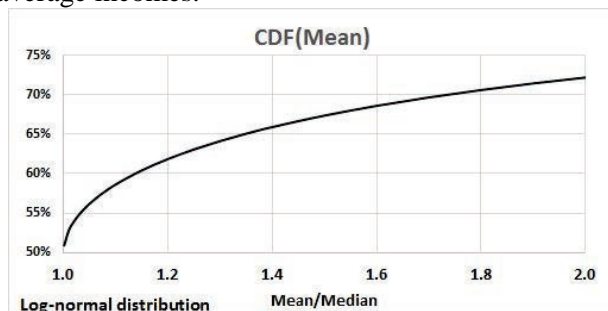

Figure 3: CDF(mean) vs Mean-Median Ratio

If Mean/Median = 1.6 (80K / 50K), then CDF(mean) = 0.6861: two-thirds of the subjects will have below-average incomes.

10. Income Shares Relative to Pro-Rata Equality

Another way to measure income inequality is to compare the income share of those with above-average incomes as a ratio of their equal share. This ratio is easily obtained for a log-normal distribution of subjects by income where all cdfs are evaluated at the mean income.

Eq. 9   Relative share for above-average incomes = (1-CDF$)/(1-CDF) = CDF/(1-CDF)
Eq. 10 Relative share for below-average incomes = CDF$/CDF = (1-CDF)/CDF

For a lognormal distribution, the relative share above always equals the inverse of the relative share below. If the mean-median ratio is 1.6 and CDF(mean) = 68%, then the relative share above equals two (67/33) while the relative share below is one-half (33/67). Those with above-average incomes have twice their equal share while those with below-average incomes have one-half of their equal or pro-rata share.

Appendix H shows the following:
Eq. 11 Average income of those making below-average income = Relative share Below * Mean Income.
Eq. 12 Average income of those making below-average income = Relative share Above * Mean Income.

If the average household income is $80,000, then the average income of those making below-average incomes is around $40,000. The average income of those making above-average incomes is $160,000.

Note that these two averages are much closer to the overall average than the incomes of those in the extremes: the top 1%, the bottom 10%, etc.

11. A Gini-Equivalent Measure of Income Equality

The Gini coefficient measures inequality without regard to the nature of the distribution. Appendix I presents the history of the Gini for US households from 1967 to 2017. It reviews the equation for the Gini coefficient given a log-normal distribution and shows it as a function of the mean-median ratio. An alternate measure, rsBelow, is proposed to measure equality. It is closely related to 1 minus the Gini coefficient.

The relative-share below (above) ratio is the below (above) average share of total income divided by their pro-rata share of income. This ratio is the ratio of (a) the average income of below (above)-average income subjects to (b) the average income of all subjects, When subjects are distributed lognormally by income the relative share below ratio has four interesting properties: (1) it is the reciprocal of the relative share above ratio, (2) it is completely determined by the mean-median ratio, (3) it is closely related to the complement of the Gini coefficient, and (4) it is a very-accessible measure of income equality.

12. Modeling Large Losses using the Log-Normal

Insurance actuaries need to model large losses. They can't use normal actuarial techniques since large losses are rare. So, they may model using the frequency with a normal distribution and the severity with a log-normal. Unfortunately there is no analytic solution for the product of these random variables. The resulting distribution is often generated using Monte-Carlo simulations.

One can model this product analytically using the product of two log-normal distributions. The normal, specified by a mean and standard deviation, can be modeled by a lognormal using the same parameters. There are two ways to do this. (1) Use Eq. 3 and Eq. 4 to determine μ and σ. (2) For a given mean, enter a slightly smaller median manually. Adjust the median up or down until the combination generates the desired standard deviation.

If $X_J \sim$ Lognormal (u_J, σ_J^2), if X_J are n *independent* log-normally distributed variables and if Y is their product, then Y is distributed lognormally: $Y \sim$ Lognormal[sum(μ_J), sum(σ_J^2)]. See Wikipedia Log-Normal. For two independent lognormal variables (frequency and severity), $Y \sim$ Lognormal ($\mu_1 + \mu_2$, $\sigma_1^2 + \sigma_2^2$).

The slight error introduced by modelling the normal with a lognormal must be compared with the uncertainty in determining the mean and standard deviation of the frequency distribution. Normally, the benefit of simplicity in analytically determining the expected loss amount that covers 95% of the situations far outweighs these small 'errors'.

13. Statistical Literacy for Decision Makers

The lognormal distribution is seldom – if ever – studied in an introductory statistics course where the focus is statistical inference. Given the interest in income inequality, the lognormal should be taught in any statistical literacy course focusing on social statistics. Here are some relevant topics and issues.

- The lognormal distribution is a "natural" distribution – both in nature and in human affairs.
- The log-normal arises naturally from the outcomes of random activities where their contributions are multiplicative. (The normal distribution arises from randomly additive activities).
- The lognormal distribution is a good fit for certain human activities involving choices: the distribution of income, the distribution of assets (wealth), the size of businesses, etc.
- The income associated with a household is the product of many factors: self-control, getting along with others, perseverance, initiative, responsibility, intelligence, ingenuity, creativity, etc.
- Seeing how a model based on randomness fits the results of human activities – many of which involve individual choices.
- The shape of the lognormal is entirely determined by the mean-median ratio. As shown in Appendix D when using the lognormal model of the 2017 CPS data, half of the total income is tied to households earning less than $120K. Recognizing this can quickly change the nature of any discussion of income inequality.
- Focus on the factors that help explain income inequality: education, number of earners, etc.
- Study the narratives involved. Many concerned about income inequality envision a bi-modal distribution for households by income: the larger mode around $30-70K; a secondary mode at a much higher level (half a million?) This narrative would certainly explain the high concentrations of income among the top 1%. But income equality arises naturally from any lognormal distribution of households by income. In a lognormal distribution, there is no secondary peak of households among the very rich. The very rich are few (top 0.1% or higher) and far between.
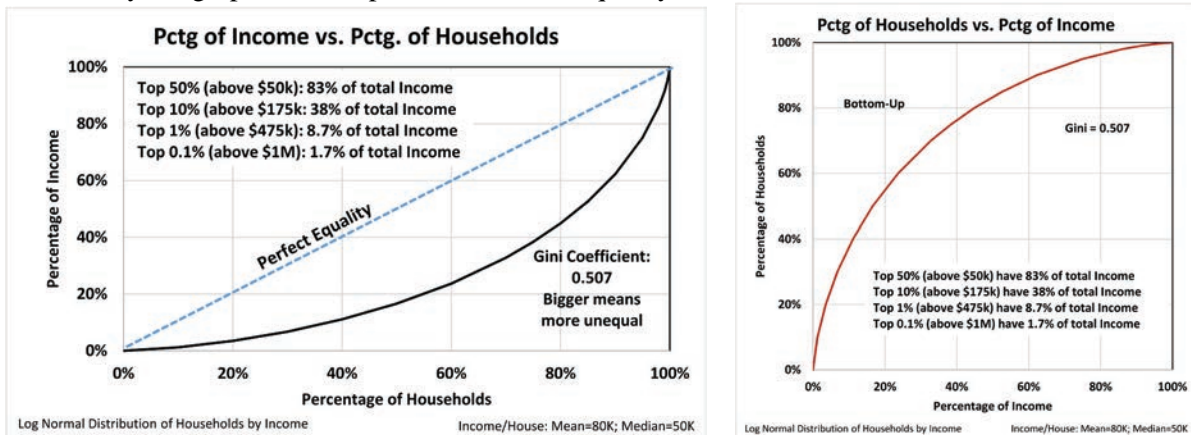- Study the graphs used to present income inequality.

Figure 4: These are Lorentz graphs. Percentage of HH and their associated percentage of total income.

The left graph is the normal presentation. The right graph is used to present the relationship as resembling a wine-glass with the rich at the very top right. Both graphs have two defects. Neither shows the change over time. They are snapshots of a given moment in time. They have no way of showing whether income inequality is increasing or decreasing. And if income inequality is increasing, they do not show by how much.
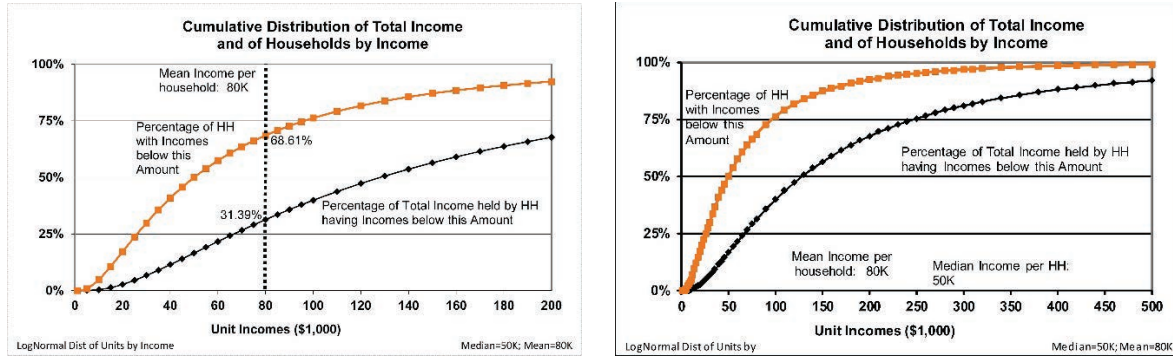
Figure 5: These are cumulative distributions of households and income. Their horizontal scale is different.

The left graph shows that at least 80% of these households earn less than $200K and they are associated with around 70% of the total income. The right graph shows the same data, but extends the horizontal scale to $500k. In both cases the difference in the percentages decreases as income increases – above the average.

14. Future Work

Work needs to be done on the following:
- Verify the parameters for the lognormal income distribution given the distribution of subjects.
- Verify the goodness of fit between the lognormal model of US income and the actual distribution.
- Obtain the mean and median incomes for those distributions using the Pareto distribution.
- Investigate the accuracy of the 40% income share estimate by Piketty (2018) for the top 1%.
- Investigate related statistical literacy issues: the difference between 'income of those in the top 1%' and 'income of those at the 99th percentile'.

15. Conclusion

This paper highlights an amazingly simple relationship between two very different percentages: If subjects are distributed lognormally by income then the share of subjects that have below-average incomes is the complement of their share of total income. This paper introduces a relative-share below (above) ratio: the below (above) average share of total income divided by their pro-rata share of income. This ratio is the ratio of (a) the average income of below (above)-average income subjects to (b) the average income of all subjects, When subjects are distributed lognormally by income the relative share below ratio has four interesting properties: (1) it is the reciprocal of the relative share above ratio, (2) it is completely determined by the mean-median ratio, (3) it is closely related to the complement of the Gini coefficient, and (4) it is a very-accessible measure of income equality. This paper argues that the lognormal distribution should be taught in any statistical literacy course for decision makers.

Acknowledgements

References

Atkinson, A. B. and T. Piketty (Ed, 2007). *Top Incomes over the Twentieth Century: A Contrast between European and English-Speaking Countries*. Oxford University Press.

Atkinson, A. B., T. Piketty and E. Saez (2011). Top Incomes in the Long Run of History. *Journal of Economic Literature* 2011, 49:1, 3–71   www.aeaweb.org/articles.php?doi=10.1257/jel.49.1.3

Aitchison, J. and Brown, J. A. C. (1957). *The Lognormal Distribution*.   Cambridge University Press. For a one-page summary, see www.komkon.org/~tacik/science/lognorm.pdf   For pages 12 and 11-113, see www.StatLit.org/pdf/1957-Aitchison-Brown-Excerpts.pdf

Congressional Budget Office (CBO 2011). Trends in the Distribution of Household Income Between 1979 and 2007.

Current Population Survey Annual Social and Economic Supplement (CPS ASEC) Research files. https://www.census.gov/data/datasets/time-series/demo/health-insurance/cps-asec-research-files.html

Dorfman, Jeffrey (2017). Higher Taxes On The Rich Are Not Enough To Stop Inequality. *Forbes*. www.forbes.com/sites/jeffreydorfman/2017/09/19/higher-taxes-on-the-rich-are-not-enough-to-stop-inequality

DQYDJ (2018). Who are the One Percent in the US by Income and Net Worth? Copy at https://dqydj.com/who-are-the-one-percent-united-states/

Efthimiou, Costas J. and Adam Wearne (2015). Household Income Distribution in the USA.

Farris, Frank (Dec 2010). The Gini Index and Measures of Inequality. *MAA Monthly 114*. P. 851-864. Copy at https://scholarcommons.scu.edu/bitstream/handle/11123/712/monthly851-864-farris.pdf

IPUMS (2017). United States Household Income Brackets and Percentiles in 2017. Copy at https://dqydj.com/united-states-household-income-brackets-percentiles/

Irfan, Mohammod (2014). Lognormal Income Inequality. Copy at www.StatLit.org/pdf/2014-Irfan-Lognormal.pdf The center of the income distribution should have been $\mu + \sigma^2$.

Krugman, Paul (2014). Wealth Over Work. The New York Times Opinion. March 24, 2014. https://www.nytimes.com/2014/03/24/opinion/krugman-wealth-over-work.html

Limpert, Stahel and Abbt (2001). Log-normal Distributions across the Sciences: Keys and Clues. *Bioscience*. Copy at www.StatLit.org/pdf/2001-Limpert-Stahel-Abbt-Bioscience2.pdf and at http://stat.ethz.ch/~stahel/lognormal/bioscience.pdf

Liu, He (2018). Proof of the Pareto Excess Mean Theorem. Personal communication.

Moll, Benjamin (2012). Lecture 6: Income and Wealth Distribution. ECO 521 Advanced MacroEcon 1. Princeton U. www.princeton.edu/~moll/ECO521Web/Lecture6_ECO521_web.pdf

Piketty, Thomas (2013 French; 2014 English). *Capital in the Twenty-First Century*. Harvard Univ. Press. Online pdf at https://dowbor.org/blog/wp-content/uploads/2014/06/14Thomas-Piketty.pdf

Piketty, Thomas, Emmanuel Saez and Gabriel Zucman (2018). Distributional National Accounts: Methods and Estimates for the United States. *The Quarterly Journal of Economics*. Vol 133, May 2018. V2. Copy at gabriel-zucman.eu/files/PSZ2018QJE.pdf

Roper, L. David (2007). Income Distribution in the United States: A Quantitative Study. Copy at http://www.roperld.com/economics/IncomeDistribution.htm

Sammartino, Frank (2017). Taxes and Income Inequality. Tax Policy Center. https://www.taxpolicycenter.org/publications/taxes-and-income-inequality/full

Schiedel, W. (2017). *The Great Leveler: Violence and the History of Inequality from the Stone Age to the 21st Century*. https://news.stanford.edu/2017/01/24/stanford-historian-uncovers-grim-correlation-violence-inequality-millennia/

Schield, M. (2018). US CPS Data for 2017 and Associated Lognormal Model Results. Copy at www.StatLit.org/Excel/2018-Schield-ASA-Data.xlsx

Schield, M. (2014). Lognormal Results for US. www.statlit.org/pdf/2014-Schield-Explore-LogNormal-Income-Excel2013.pdf

Schield, M. (2015). Lognormal spreadsheet. www.StatLit.org/Excel/2015-Schield-Lognormals-Ratio.xlsx

Sommeiller, E., M. Price and E Wazeter (2016). *Income inequality in the U.S. by state, metropolitan area, and county*. Economic Policy Institute. https://www.epi.org/publication/income-inequality-in-the-us/

Sommeiller, E & M. Price (2018). *The new gilded age: Income inequality in the U.S.* Economic Policy I. www.epi.org/publication/the-new-gilded-age-income-inequality-in-the-u-s-by-state-metropolitan-area-and-county

Stone C., D. Trisi, A. Sherman, R. Taylor. A guide to Statistics on Historical Trends in Income Inequality. www.cbpp.org/research/poverty-and-inequality/a-guide-to-statistics-on-historical-trends-in-income-inequality

US Census Current Population Survey (2017): HINC-06. Income Distribution to $250,000. https://www.census.gov/data/tables/time-series/demo/income-poverty/cps-hinc/hinc-06.2017.html

US Census Statistical Abstract (1998). Table 741.

Wilkinson, R. and K. Pickett (2010). *The Spirit Level: Why Equality is Better for Everyone*. Penguin books.

Wikipedia Household income US: https://en.wikipedia.org/wiki/household_income_in_the_United_States

Appendix A: Modelling Incomes: Methodology

Sommeiller and Price (2018, Appendix A) described their methodology as follows:

> *Knowing the amount of income and the number of taxpayers in each bracket, we can use the properties of a statistical distribution known as the Pareto distribution to extract estimates of incomes at specific points in the distribution of income, including the 90th, 95th, and 99th percentiles.[31] With these threshold values we then calculate the average income of taxpayers with incomes that lie between these ranges, such as the average income of taxpayers with incomes greater than the 99th percentile (i.e., the average income of the top 1 percent).*

> *Pareto Interpolation:*
> *In a study of the distribution of incomes in various countries, the Italian economist Vilfredo Pareto observed that as the amount of income doubles, the number of people earning that amount falls by a constant factor. In the theoretical literature, this constant factor is usually called the Pareto coefficient (labeled* bi *in **Appendix Table A5**).[42] Combining this property of the distribution of incomes with published tax data on the number of tax units and the amount of income at certain levels, it is possible to estimate the top decile (or the highest-earning top 10 percent of tax units), and within the top decile, a series of percentiles such as the average annual income earned by the highest-income 1 percent of tax units, up to and including the top 0.01 percent fractile (i.e., the average annual income earned by the richest 1 percent of the top 1 percent of tax units).*
> *42. See Atkinson and Piketty (2007) for a discussion of Pareto interpolation.*

Atkinson, Piketty and Saez (2011, p. 13) described their methodology as follows:

> *The key property of Pareto distributions is that the ratio of average income y*(y) of individuals with income above y to y does not depend on the income threshold y.*
> *ie, y*(y)/y = β with β=α/(α−1).*
> *That is, if β = 2, the average income of individuals with income above $100,000 is $200,000 and the average income of individuals with income above $1 million is $2 million. Intuitively, a higher β means a fatter upper tail of the distribution. From now on, we refer to β as the inverted Pareto coefficient. Throughout this paper, we choose to focus on the inverted Pareto coefficient β (which has more intuitive economic appeal) rather than the standard Pareto coefficient α. Note that there exists a one-to-one, monotonically decreasing relationship between the α and β coefficients, i.e., β = α/(α − 1) and α = β/(β − 1) (see table 3).*

Atkinson, Piketty and Saez (2007, p. 10) described the impact of the top shares on overall inequality:

> *It might be thought that top shares have little impact on overall inequality. If we draw a Lorenz curve, defined as the share of total income accruing to those below percentile p, as p goes from 0 (bottom of the distribution) to 100 (top of the distribution), then the top 1 percent would scarcely be distinguishable on the horizontal axis from the vertical endpoint, and the top 0.1 percent even less so. The most commonly used summary measure of overall inequality, the Gini coefficient, is more sensitive to transfers at the center of the distribution than at the tails. (The Gini coefficient is defined as the ratio of the area between the Lorenz curve and the line of equality over the total area under the line of equality.)*

> *But top shares can materially affect overall inequality, as may be seen from the following calculation. If we treat the very top group as infinitesimal in numbers, but with a finite share S* of total income, then, graphically, the Lorenz curve reaches 1 − S* just below p = 100. As a result, the total Gini coefficient can be approximated by S* + (1 − S*) G, where G is the Gini coefficient for the population excluding the top group (Atkinson 2007b). This means that, if the Gini coefficient for the rest of the population is 40 percent, then a rise of 14 percentage points in the top share, as happened with the share of the top 1 percent in the United States from 1976 to 2006, causes a rise of 8.4 percentage points in the overall Gini. This is larger than the official Gini increase from 39.8*

percent to 47.0 per- cent over the 1976–2006 period based on U.S. household income in the Current Population Survey (U.S. Census Bureau 2008, table A3)

Piketty (2007, p. 260), described Pareto's motivation and the Pareto family of distributions:

Pareto's judgment was clearly influenced by his political prejudices: he was above all wary of socialists and what he took to be their redistributive illusions. In this respect he was hardly different from any number of contemporary colleagues, such as the French economist Pierre Leroy-Beaulieu, whom he admired. Pareto's case is interesting because it illustrates the powerful illusion of eternal stability, to which the uncritical use of mathematics in the social sciences sometimes leads. Seeking to find out how rapidly the number of taxpayers decreases as one climbs higher in the income hierarchy, Pareto discovered that the rate of decrease could be approximated by a mathematical law that subsequently became known as "Pareto's law" or, alternatively, as an instance of a general class of functions known as "power laws." Indeed, this family of functions is still used today to study distributions of wealth and income. Note, however, that the power law applies only to the upper tail of these distributions and that the relation is only approximate and locally valid. It can nevertheless be used to model processes due to multiplicative shocks, like those described earlier.

Piketty (2007, p. 418-421), described the nature and use of the Pareto coefficient:

19.… The Pareto coefficient, about which I will say more in Chapter 10, enables us to relate the shares of the top decile, top centile, and top thousandth:

32. The simplest way to think of Pareto coefficients is to use what are sometimes called "inverted coefficients," which in practice vary from 1.5 to 3.5. An inverted coefficient of 1.5 means that average income or wealth above a certain threshold is equal to 1.5 times the threshold level (individuals with more than a million euros of property own on average 1.5 million euros' worth, etc., for any given threshold), which is a relatively low level of inequality (there are few very wealthy individuals). By contrast, an inverted coefficient of 3.5 represents a very high level of inequality. Another way to think about power functions is the following: a coefficient around 1.5 means that the top 0.1 percent are barely twice as rich on average as the top 1 percent (and similarly for the top 0.01 percent within the top 0.1 percent, etc.). By contrast, a coefficient around 3.5 means that they are more than five times as rich. .. explained in the online technical appendix.

Moll (2012) indicated the proof of the unique ratio for a Pareto distribution:

In practice, quite often [we] don't have data on share of top 1%

• Use Pareto interpolation. Assume income has CDF: $F(y) = 1 – (y/k)^{-alpha}$

• Useful property of Pareto distribution: average above threshold proportional to threshold

$$E[\bar{y} \mid \bar{y} \geq y] = [Integral\ from\ y\ to\ \infty\ of\ z*f(z)dz]\ divided\ by\ 1-F(y) = y[alpha / (alpha-1)]$$

• Estimate and report Beta = alpha / (alpha – 1)

• Example: Beta = 2 means average income of individuals with income above $100,000 is $200,000 and average income of individuals with income above $1 million is $2 million.

• Obviously imperfect, but useful because alpha or beta is exactly what our theories generate.

To convert the mean excess income into a percentage of the total income for the top X percent, multiply by X percent and divide by the average income. Consider an income distribution having an average of $50,000 where the high-end best-fit Pareto has alpha = 2 (Beta = 2).

- If the threshold for the top 10% is $100,000, then the average income for those subjects earning more than $100,000 is $200,000. The fraction of the total income earned by this top 10% is 40%: four times their equal share. Income share = 0.1 * $200,000 / $50,000.
- If the threshold for the top 1% is $250,000, then the average income for those subjects earning more than $250,000 is $500,000. The fraction of the total income earned by this top 1% is 10%: ten times their equal share. Income share = 0.01 * $500,000 / $50,000.

For more background, search on 'mean excess theorem', 'mean excess function' and 'mean excess plot'.

Wilkenson and Pikett (2009) described their use of statistical distributions:

> Statistical distribution: We investigated two statistical distributions commonly used to model wealth in populations: the Pareto and the log–normal. In the modelling of wealth, each of these distributions is defined by the mean wealth per capita – as the central tendency – and the Gini coefficient – as a measure of dispersion. Each distribution has a characteristic shape (Fig. 1). The logarithm of the Pareto distribution is heavily right-skewed, with a sharp cut-off on the left side. The logarithm of the log–normal distribution is a symmetric normal distribution, with no skew. By combining estimates from both of these distributions, we could therefore examine wealth distributions with varying degrees of skew.

Piketty, Saez and Zucman (2018). This article attempts to compute inequality statistics for the United States that overcome the limits of existing series by creating distributional national accounts. We combine tax, survey, and national accounts data to build new series on the distribution of national income since 1913. In contrast to previous attempts that capture less than 60% of U.S. national income— such as Census Bureau estimates (U.S. Census Bureau 2016) and top income shares (Piketty and Saez 2003)—our estimates capture 100% of the national income recorded in the national accounts.

Appendix B: The Pareto Distribution

The Pareto distribution models random processes above a positive minimum value. Consider $sP(x) = Xmin/(1-U)$ where U is a positive uniform random variable that is less than one. This pseudo-Pareto generates random values that are greater than Xmin with decreasing density. This pseudo Pareto is less flexible than the actual Pareto: $P(x) = Xmin/[(1-U)^{(1/\gamma)}]$ where $\gamma > 1$ and U is a uniform random variable with values between zero and one. Both distributions avoid values of infinity.

The Pareto has a unique mathematical property. Moll noted that the average above a threshold is proportional to the value of the threshold.

Eq. 13 *Assume income has CDF: F(y) = 1 – (y/k)^{-alpha}* [where alpha is less than one]

Eq. 14 $E[\bar{y} / \bar{y} \geq ymin] =$ *[Integral from ymin to ∞ of y*f(y)dy] divided by 1-F(ymin)*

Eq. 15 $E[\bar{y} / \bar{y} \geq ymin] = ymin *[alpha / (alpha-1)] = y * Beta$

Liu (2018) provided this proof of the Pareto mean excess formula:

Eq. 16 Let F(y) = cumulative distribution function = $1 – [(y/Ymin)^{(-alpha)}]$ with alpha > 1 and y ≥ Ymin.

Eq. 17 F'(y) = the Pareto probability distribution function = $alpha*[Ymin^{alpha}]*[y^{(-alpha - 1)}]$.

Eq. 18 y*F'(y) = the probability distribution function of income = $alpha*[Xmin^{alpha}]*[y^{(-alpha)}]$.

Eq. 19 E(y|y>Ymin) = ave. income above Ymin = integral of yF'(y) from Ymin to infinity/[1-F(Ymin)].

Eq. 20 E(y|y>Ymin) = {[alpha/(1-alpha)]*y} evaluated at infinity minus Ymin.

The evaluation at these limits reverses the sign and replaces y with Ymin in the preceding equation.

Eq. 21 E(y|y>Ymin) = Pareto mean excess income = [alpha/(alpha-1)]*[Ymin].

For alpha = 2, the average income above Ymin is double Ymin.

Piketty (2017) noted that 'the share of income above any cutoff' divided by 'the fraction of subjects having incomes above this cutoff' was determined by the Inverse Pareto coefficient: a property of the Pareto distribution. If this ratio was 5, then those above this cutoff had five times their equal share.

Appendix C Generating Lognormal Underlying Parameters from External Statistics

Source equations from Wikipedia: Log-Normal Distribution

Eq. 22 Median = Exp(μ)
Eq. 23 Mean = Exp($\mu + \sigma^2/2$)
Eq. 24 Variance = StdDev^2 = [Exp(σ^2) -1]*Exp($2\mu + \sigma^2$)

Solving for μ and σ in terms of the mean and median gives:

Eq. 25 μ = Ln(median)		From Eq. 22
Eq. 26 Mean/Median = Exp($\mu + \sigma^2/2$) / Exp(μ)		From Eq. 22 and Eq. 23
Eq. 27 Ln(Mean / Median) = ($\mu + \sigma^2/2$) - μ		From Eq. 26
Eq. 28 σ^2 = 2*Ln(mean/median) = Ln[(Mean/Median)^2]		From Eq. 27
Eq. 29 σ = Sqrt[2*Ln(mean/median)] = Sqrt{Ln[(Mean/Median)^2]}		From Eq. 28

Solving for μ and σ in terms of the mean and standard deviation gives:

Eq. 30 $\mu + \sigma^2/2$ = Ln(Mean)	From Eq. 23
Eq. 31 Var = [exp(σ^2) - 1]*exp($2*\mu + \sigma^2$)	Eq. 24
Eq. 32 Mean = exp($\mu + \sigma^2/2$)	From Eq. 30
Eq. 33 Var = [exp(σ^2) - 1]*Mean^2	From Eq. 31 and Eq. 32
Eq. 34 Coef of Variation = StdDev/Mean = Sqrt(Variance)/Mean	Definitions
Eq. 35 StdDev/Mean = Sqrt[Exp(σ^2) - 1]	From Eq. 33
Eq. 36 σ^2 = Ln[(StdDev/Mean)^2 + 1]	From Eq. 35
Eq. 37 σ = Sqrt{Ln[(StdDev/Mean)^2 + 1]}	From Eq. 36
Eq. 38 μ = Ln(Mean) – σ^2	From Eq. 30
Eq. 39 μ = Ln(Mean) – Ln[(StdDev/Mean)^2 + 1]	From Eq. 38

Solving for μ and σ in terms of the median and standard deviation gives:

Eq. 40 μ = Ln(Median)	From Eq. 25
Eq. 41 $\mu + \sigma^2/2$ = Ln(Mean)	From Eq. 30
Eq. 42 Ln(Mean) = Ln(Median) + $\sigma^2/2$.	From Eq. 41
Eq. 43 Variance = [Exp(σ^2) -1]*Exp($2\mu + \sigma^2$)	Eq. 24
Eq. 44 Variance = [Exp(σ^2) -1]*Exp($2*$Ln(Median) + σ^2)	From Eq. 43

A simple analytic solution for sigma is not obvious in terms of the median and standard deviation. An alternate approach is trial-and-error. Try different values of sigma to see which generates the desired standard deviation. Similarly, a simple analytic solution for sigma is not obvious in terms of the Gini coefficient. Again, the alternative is to use trial-and-error.

Appendix D Lognormal Model for 2017 Census CPS Results

The Census CPS income is regular (repeatable) income. It omits capital gains. According to the 2017 Census CPS, the mean income for the sampled households is $81,700. The median of the grouped data is obtained by linear interpolation: $67,100. The associated lognormal generates this

A	B	C	D	E	F	G	H	I	
	----------BOTTOM-UP----------			---TOP_DOWN---		Times=Share:	Ave$	Ave$	2
%#Up	$Cutoff#	%$Up		%#down	%$down	%$TD/%#TD	Above	/$Cutoff#	3
0%	0.0	0.00%		100%	100.0%	1.0	86		4
10%	21.6	1.79%		90%	98.2%	1.1	94	4.3	5
20%	31.0	4.85%		80%	95.1%	1.2	103	3.3	6
30%	40.2	8.98%		70%	91.0%	1.3	112	2.8	7
40%	50.2	14.21%		60%	85.8%	1.4	123	2.5	8
50%	61.7	20.67%		50%	79.3%	1.6	137	2.2	9
60%	75.9	28.62%		40%	71.4%	1.8	154	2.0	10
70%	94.7	38.46%		30%	61.5%	2.1	177	1.9	11
75%	107.1	44.30%		25%	55.7%	2.2	192	1.8	12
80%	122.8	50.95%		20%	49.0%	2.5	211	1.7	13
85%	144.0	58.65%		15%	41.3%	2.8	238	1.6	14
90%	176.0	67.86%		10%	32.1%	3.2	277	1.6	15
95%	236.8	79.59%		5%	20.4%	4.1	352	1.5	16
98%	330.9	89.18%		2%	10.8%	5.4	466	1.4	17
99%	413.5	93.43%		1%	6.6%	6.6	566	1.4	18
99.5%	507.1	96.06%		0.5%	3.9%	7.9	679	1.3	19
99.9%	772.3	98.85%		0.1%	1.2%	11.5	994	1.3	20
99.95%	909.8	99.33%		0.05%	0.7%	13.4	1,156	1.3	21
99.99%	1,291.6	99.81%		0.01%	0.2%	18.6	1,602	1.2	22

CELL	FormulaText()			CELL	FormulaText()	
A4:A22	Already entered and B4:C4. I5 is empty			F4	=1-C4	
B5	=LOGNORM.INV(A5,T$4,T$7)			G4	=F4/E4	
C5	=LOGNORM.DIST(B5,T$17,T$18,1)			H4	=O$4*F4/E4	
E4	=1-A4			I5	=H5/B5	

Table 4: Log-Normal Distribution of US households and income by CPS household income

Appendix E presents the Aitchison and Brown (1957) graphical argument using the Lorenz curve. Appendix F presents an algebraic argument similar to the one given by Aitchison and Brown (1957). Appendix G presents a graphical proof by the author using the underlying normal distributions.

The percentage of subjects that have below-average incomes is a simple way of measuring income inequality. For a lognormal distribution, this percentage is identical to the percentage of total income that is associated with subjects having above-average incomes. The higher these percentages, the greater the income inequality. They may be overly simplistic, but they are readily understandable: much more so than the Gini coefficient.

If the mean-median ratio is 1.6, two-thirds of the subjects have below-average incomes (two-thirds of the total income is associated with subjects having above-average incomes).

Appendix E: Graphical Proof #1 of the Lognormal Balance

Figure 4 presents the Lorenz curve for the lognormal distribution of subjects and the lognormal distribution of incomes. The horizontal axis is the cumulative distribution function (cdf) for the lognormal distribution of subjects by income; the vertical axis is the cumulative distribution function (cdf) for the associated lognormal distribution of incomes.

Aitchison and Brown (1957, p. 112) noted that if the average income on the horizontal axis intersects the Lorenz curve at the 45 degree line then the CDF(mean) must equal one minus CDF$(mean).
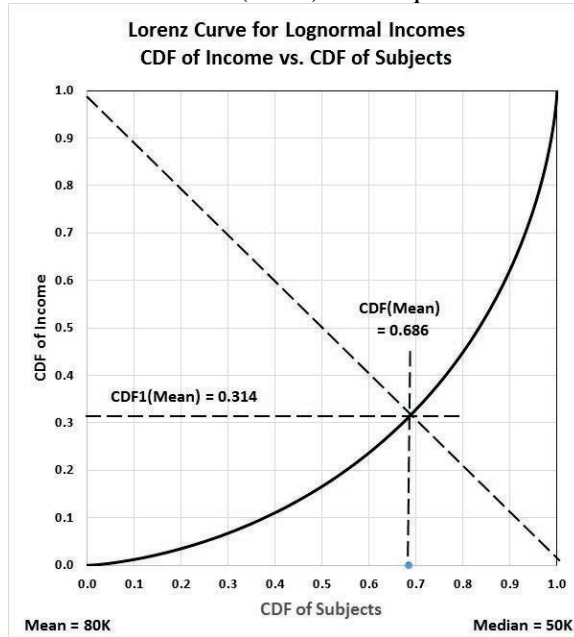

Figure 6: Lorenz Curve of Income vs Subjects

Unfortunately, the argument supporting the truth of their premise is somewhat obscure.

Appendix F: Algebraic Proof of the Lognormal Balance

Assume that subjects are distributed lognormally by income. Mean and median apply to the distribution of subjects by income. CDF$(mean) is the percentage of total income that is earned by subjects with below-average incomes; 1-CDF$(mean) is the percentage of total income that is earned by subjects with above-average incomes. This derivation follows the derivation provided by Aitchison and Brown (1957).

Consider the fraction of Subjects that have below-average income.

Eq. 45 $CDF(mean) = Phi[(Ln(mean) – \mu)/\sigma]$

Eq. 46 $CDF(mean) = Phi\{Sqrt[Ln(mean/median)/2]\}$

Consider the fraction of Total Income that is earned by subjects with above-average income.

Eq. 47 $1 – CDF\$(mean) = 1 – Phi[(Ln(mean)–\mu1)/\sigma]$

Eq. 48 $1 – CDF\$(mean) = Phi[(\mu1 - Ln(mean))/\sigma]$

Eq. 49 $1 – CDF\$(mean) = Phi[(\mu + \sigma2 - Ln(mean))/\sigma]$

Eq. 50 $\mu + \sigma2 = Ln(median) + 2*Ln(mean) – 2*Ln(median)$

Eq. 51 $1 – CDF\$(mean) = Phi[Ln(mean/median)/\sigma]$

Eq. 52 $1 – CDF\$(mean) = Phi\{Sqrt[Ln(mean/median)/2]\}$

Note: Eq. 52 = Eq. 46.   QED.   Eq. 48 is critical: $1 - CDF(Z) = CDF(-Z)$.

Appendix G: Graphical Proof #2 of the Lognormal Balance

The following is an alternate graphical proof of the log-normal balance.

Figure 5 shows the two underlying Normal distributions. They have the same standard deviations but have centers that are separated by σ^2. These two Normal curves must intersect half-way between their centers at $\mu + \sigma^2/2$. This point of intersection is always the log-normal mean of the left (subject) distribution.
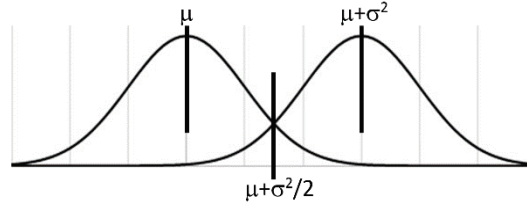

Figure 7: Income Balance with the Underlying Normal Distributions

Since these two Normal distributions have identical standard deviations, the area to the left of the intersection point under the left Normal (the subject Normal) must equal the area to the right of the intersection point under the right Normal (the income Normal). The balance condition states this equality of areas in words.

Appendix H: Relative Share of Income

Another way to describe income inequality is to say that the subjects with above-average incomes have X times their equal (pro-rata) share. The following is a derivation of this ratio. Assume that all the CDF are being evaluated at the mean income for the entire population. Thus, CDF = CDF(mean)

Define rsBelow (rsAbove) as the relative share of total income for subjects having below-average (above-average) incomes.

Eq. 53   rsBelow = CDF$ / CDF = [1-CDF]/CDF for a lognormal dist.
Eq. 54   rsAbove = [1-CDF$] / [1-CDF] = CDF/[1-CDF] for a lognormal.

In words, rsBelow is the ratio of (1) the percentage of total income that is earned by subjects having below-average incomes, to (2) the percentage of subjects that have below-average incomes; rsAbove is the share of total income that is held by those with above-average incomes relative to their equal or pro-rata share.

These ratios are directly related to two averages: the mean income among subjects with below-average income (meanBelow) and the mean income among subjects with above-average income (meanAbove).

Eq. 55   meanBelow = CDF$*TotIncome / [CDF*TotNumber] = Mean * CDF$/CDF
Eq. 56   meanAbove = [1-CDF$]*TotIncome/[(1-CDF)*Tot#] = Mean * (1-CDF$)/(1-CDF)

For a log-normal distribution, this gives
Eq. 57 meanBelow = [1-CDF]*Mean / CDF = Mean*rsBelow
Eq. 58 meanAbove = CDF*Mean/[1-CDF] = Mean*rsAbove

For a lognormal distribution, both relative-share ratios are determined by MMR: the mean-median ratio:

Eq. 59 rsBelow = [1-Phi{Sqrt[Ln(MMR)/2]}]/Phi{Sqrt[Ln(MMR)/2]}
Eq. 60 rsAbove = Phi{Sqrt[Ln(MMR)/2]}/[1 - Phi{Sqrt[Ln(MMR)/2]}]

Appendix I: The Gini Coefficient

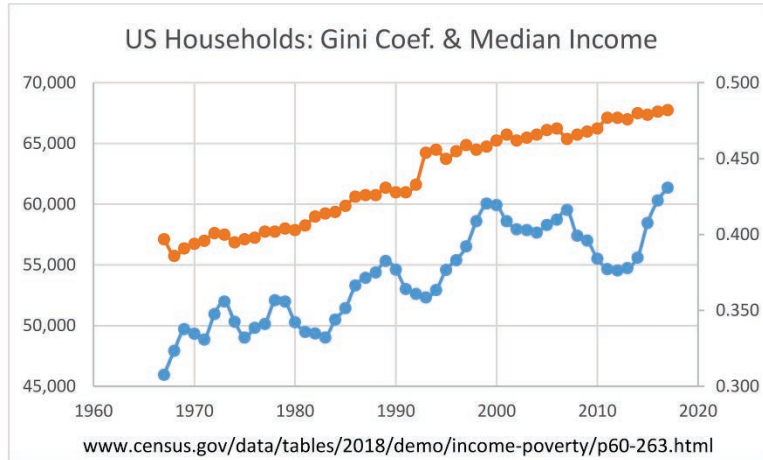The Gini coefficient measures inequality of outcomes.

Figure 8: Gini Coefficient and Median Income for US Households: 1968 to 1917.

Aitchison and Brown (1957) showed that for a lognormal distribution, the Gini coefficient is

Eq. 61   Gini = Erf(σ/2) = 2 Phi[σ/Sqrt(2)] - 1

This can be restated in terms of the mean-median ratio:

Eq. 62   Gini = 2 Phi{Sqrt[Ln(mean/median)]} - 1

Farris (2010) related the Gini coefficient to P where P is:

Eq. 63   P = (Gini+1)/2

If subjects are distributed lognormally, P is:

Eq. 64   P = Phi{Sqrt[Ln(Mean/Median)]}

Note that Eq. 23 and Eq. 8 differ just slightly but P will always exceed CDF(mean) since Phi(X) > Phi[X/Sqrt(2)] for X > 0.

Farris described P as the average or expected income percentile. Note that P is not the same as the percentile of the average income.

Unfortunately both P and Gini are somewhat inaccessible and they both measure inequality. If equality is considered the good, then an alternate approach is to look for a measure of equality.

This paper proposes rsBelow, the ratio of meanBelow income to average income, as a simple, accessible measure of equality. Figure 5 shows the close relationship between this measure of equality and the complement of the Gini coefficient.
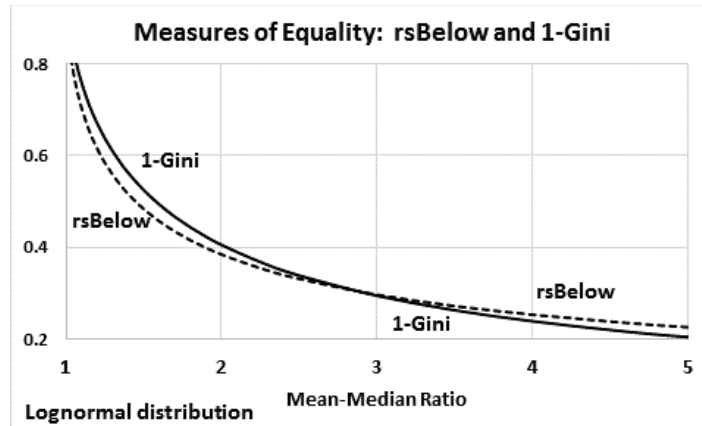
Figure 9: Measures of Lognormal Equality: rsBelow and 1-Gini

Note that the 1-Gini and rsBelow cross at mean/median = 2.824 where they equal 0.308. For smaller mean-median ratios, rsBelow is conservative: it underestimates the amount of equality. When mean/median = 1.15, rsBelow is most conservative: 5.4 points below 1-Gini.

If mean/median is 1.6, then the fraction of subjects that have below-average incomes is 0.686, P is 0.754, rsBelow is 0.458 and 1-Gini is 0.493. When rsBelow is 0.458, those with below-average incomes have 45.8% of their equal share of income while those with above-average incomes have 2.18 times their pro-rata share.

Appendix J: Comparing the Lognormal and the Normal

Limpert, Stahel and Abbt (2001) made the following comparisons of the normal and the lognormal:
Why the normal distribution is so popular.

> Regardless of statistical considerations, there are a number of reasons why the normal distribution is much better known than the log-normal. A major one appears to be symmetry, one of the basic principles realized in nature as well as in our culture and thinking. Thus, probability distributions based on symmetry may have more inherent appeal than skewed ones. Two other reasons relate to simplicity. First, as Aitchison and Brown (1957, p. 2) stated, "Man has found addition an easier operation than multiplication, and so it is not surprising that an additive law of errors was the first to be formulated." Second, the established, concise description of a normal sample—$\bar{x} \pm s$—is handy, well-known, and sufficient to represent the underlying distribution, which made it easier, until now, to handle normal distributions than to work with log-normal distributions. Another reason relates to the history of the distributions: The normal distribution has been known and applied more than twice as long as its log-normal sister distribution. Finally, the very notion of "normal" conjures more positive associations for non-statisticians than does "log-normal." For all of these reasons, the normal or Gaussian distribution is far more familiar than the log-normal distribution is to most people.

Conclusion:
On the first page of their book, Aitchison and Brown (1957) stated that, compared with its sister distributions, the normal and the binomial, the log-normal distribution "has remained the Cinderella of distributions, the interest of writers in the learned journals being curiously sporadic and that of the authors of statistical textbooks but faintly aroused." This is indeed true: Despite abundant, increasing evidence that lognormal distributions are widespread in the physical, biological, and social sciences, and in economics, log-normal knowledge has remained dispersed. The question now is this: Can we begin to bring the wealth of knowledge we have on normal and log-normal distributions to the public? We feel that doing so would

lead to a general preference for the lognormal, or multiplicative normal, distribution over the Gaussian distribution when describing original data.

Appendix K: Results for a particular Log-Normal Distribution

In round numbers the US average and median incomes for the distribution of households by income are $50,000 and $80,000 respectively. Table 6 shows the associated data for the lognormal distribution of subjects and the lognormal distribution of incomes by household income.

%#Up	-----BOTTOM-UP------			---TOP-DOWN---		Times=Share: Down:%$ / %#
	$Cutoff#	%$Up		%#down	%$down	
0%	0.0	0.00%		100%	100.0%	1.0
10%	14.4	1.22%		90%	98.8%	1.1
20%	22.1	3.51%		80%	96.5%	1.2
30%	30.1	6.76%		70%	93.2%	1.3
40%	39.1	11.07%		60%	88.9%	1.5
50%	50.0	16.61%		50%	83.4%	1.7
60%	63.9	23.69%		40%	76.3%	1.9
70%	83.1	32.81%		30%	67.2%	2.2
75%	96.2	38.40%		25%	61.6%	2.5
80%	113.1	44.91%		20%	55.1%	2.8
85%	136.6	52.67%		15%	47.3%	3.2
90%	173.2	62.25%		10%	37.8%	3.8
95%	246.4	75.03%		5%	25.0%	5.0
98%	366.2	86.09%		2%	13.9%	7.0
99%	477.0	91.26%		1%	8.7%	8.7
99.5%	607.5	94.59%		0.5%	5.4%	10.8
99.9%	1,000.4	98.30%		0.1%	1.7%	17.0
99.95%	1,214.8	98.99%		0.05%	1.0%	20.3
99.99%	1,840.4	99.70%		0.01%	0.3%	29.8

Table 5: Distribution of US households and income by household income

Those that see income inequality as a big problem focus on the far-right column: the super-rich have many times their equal share. The top 0.01% of households have incomes that are almost 30 times their equal share.

Those that see income inequality as a small problem focus on the second column from the right. Collectively the total income of the top 0.01% of households is only 0.3% of the total.

The super-rich households are high severity – but low frequency. Suppose the government confiscates all the income from the top 0.01% of households and distributes it to the households with below-average incomes (less than $80K). The increase would be small: less than 1% (0.3% of total income on top of 31.4% of total income).

Source: www.statlit.org/pdf/2014-Schield-LogNormal-Income2B-Excel2013-Demo.pdf