Statistical Literacy: The Lognormal Distribution

Milo Schield, Augsburg U.
Editor: www.StatLit.org
US Rep: International Statistical Literacy Project

Amer. Statistical Association JSM
July 30, 2018
www.StatLit.org/
pdf/2018-Schield-ASA-Slides.pdf
XLS/Explore-LogNormal-Incomes-Excel2013.xlsx

Best-selling statistics books

80 million: *World Almanac (Since 1896)*

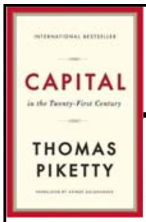
5 million: Economist: *World in Figures* (200K/yr; 25 years)

1.5 million: Piketty (2017): *Capital in the 21st Century*

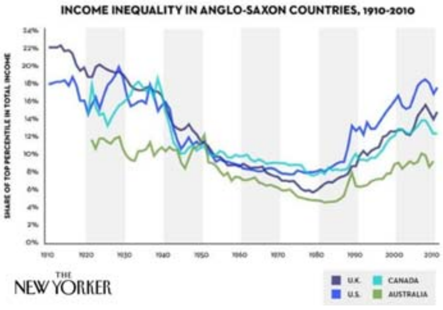
500,000: Murray & Hernstein (1994): *The Bell Curve*

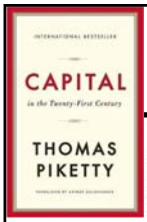
200,000: Hacker (1992): *"Two Nations: Black and White, Separate, Hostile, Unequal."*

https://www.washingtonpost.com/archive/lifestyle/1995/09/22/black-and-white-read-all-over-the-hot-books-that-make-the-melting-pot-boil/ee1de9b5-a172-4dfd-bb7a-1eb1d6cf9d77/

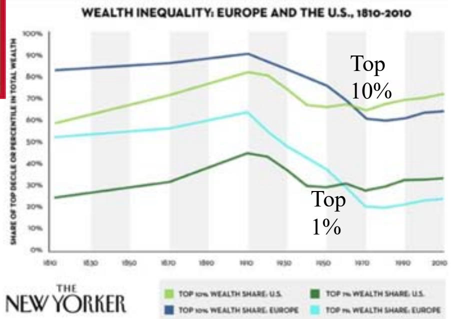


Capital in the 21st Century: Income by Country (Top 1%)





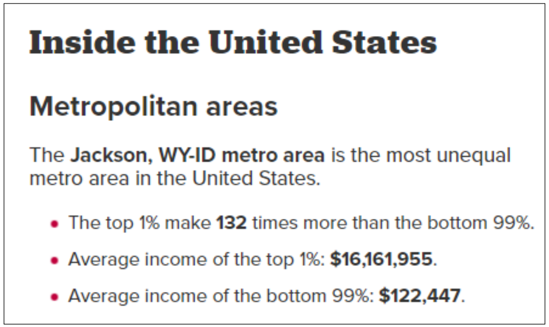
Capital in the 21st Century: Wealth by Country (Top 1%)



Top 10%

Top 1%

EPI (2018): US Income Inequality by Metro Area



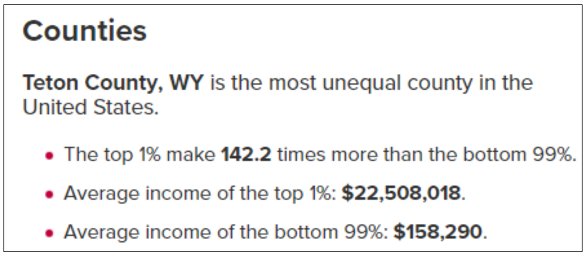
Inside the United States

Metropolitan areas

The **Jackson, WY-ID metro area** is the most unequal metro area in the United States.

- The top 1% make **132** times more than the bottom 99%.
- Average income of the top 1%: **$16,161,955**.
- Average income of the bottom 99%: **$122,447**.

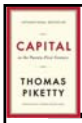
EPI (2018): US Income Inequality by County



Counties

Teton County, WY is the most unequal county in the United States.

- The top 1% make **142.2** times more than the bottom 99%.
- Average income of the top 1%: **$22,508,018**.
- Average income of the bottom 99%: **$158,290**.



Piketty: Censored Data Problem

Evaluate income share held by top 1% over time.

Data source: Tax data

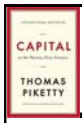
Problem: Tax authorities censors high-income data.

So, how did Piketty deduce the income share of top 1%

Piketty used a model: the Pareto distribution.

By fitting this model to uncensored incomes, he inferred the distribution of the censored incomes.

Atkinson et al (2011). P 12-14.



Piketty: *Capital in the 21^{st} Century*

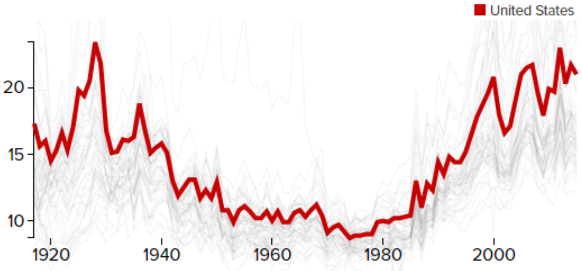
The key property of Pareto distributions is this: the "ratio of 'average income y*(y) of individuals with income above y' to y does not depend on the income threshold y."

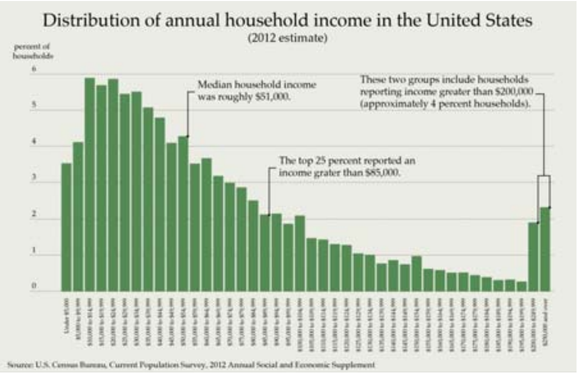
[Ave Income > y] / y = Beta

"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."

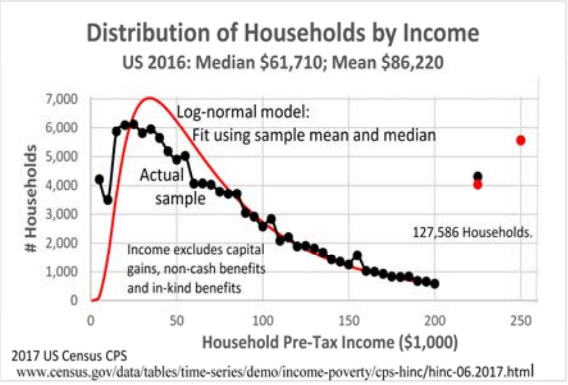
Atkinson, Piketty, Suan (2011). P 12-14.

EPI (2018): Income Inequality Top 1% Since 1920 by Country




Distribution of annual household income in the United States
(2012 estimate)

Source: U.S. Census Bureau, Current Population Survey, 2012 Annual Social and Economic Supplement


Distribution of Households by Income
US 2016: Median $61,710; Mean $86,220

2017 US Census CPS
www.census.gov/data/tables/time-series/demo/income-poverty/cps-hinc/hinc-06.2017.html

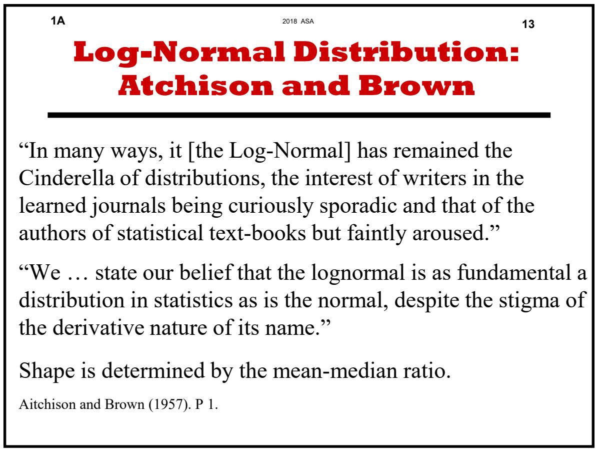
Log-Normal Distribution

Log-Normal shape is common. Examples:

- Incomes (bottom 97%), assets, size of cities
- Weight and blood pressure of humans (by gender)
- Stock and portfolio returns

Log-Normal is useful.

- Function is easier to work with than a histogram
- Understand what determines or explains shape
- calculate the share of total income held by the top X%
- calculate share of total income held by the 'above-average'
- explore effects of change in mean-median ratio.



1A 2018 ASA 13

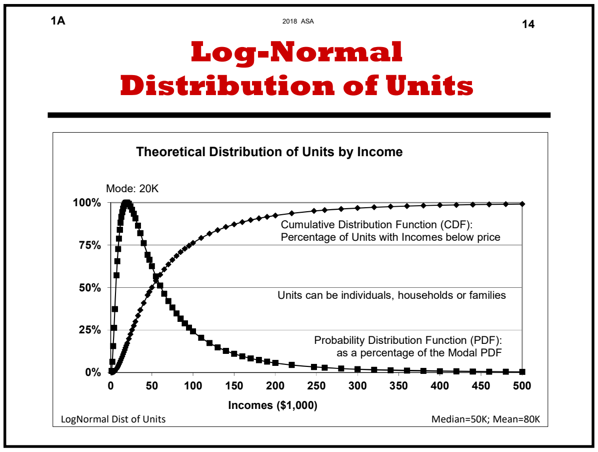
Log-Normal Distribution: Atchison and Brown

"In many ways, it [the Log-Normal] has remained the Cinderella of distributions, the interest of writers in the learned journals being curiously sporadic and that of the authors of statistical text-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."

Shape is determined by the mean-median ratio.

Aitchison and Brown (1957). P 1.

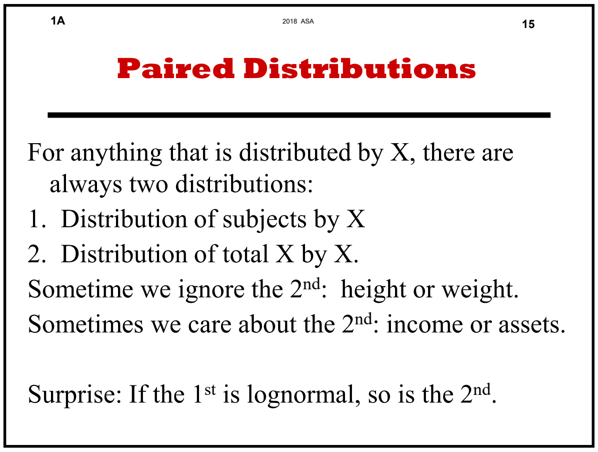


1A 2018 ASA 14

Log-Normal Distribution of Units

Theoretical Distribution of Units by Income

LogNormal Dist of Units Median=50K; Mean=80K



1A 2018 ASA 15

Paired Distributions

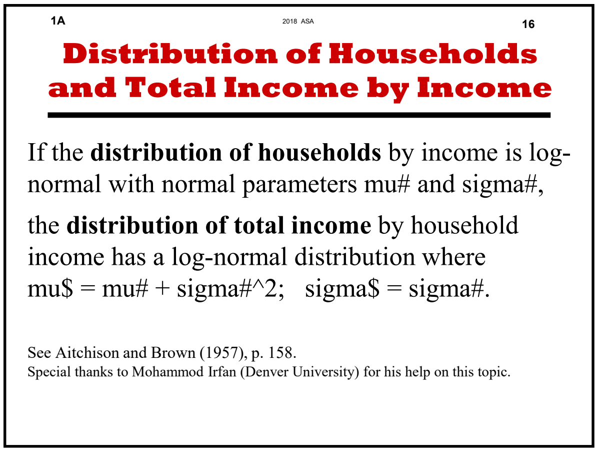
For anything that is distributed by X, there are always two distributions:

1. Distribution of subjects by X
2. Distribution of total X by X.

Sometime we ignore the 2nd: height or weight.

Sometimes we care about the 2nd: income or assets.

Surprise: If the 1st is lognormal, so is the 2nd.



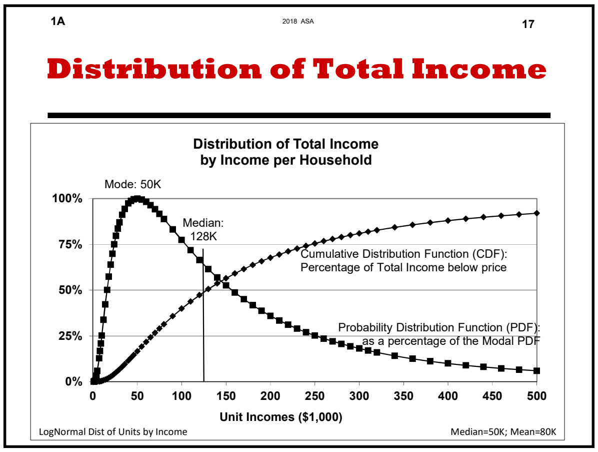
1A 2018 ASA 16

Distribution of Households and Total Income by Income

If the **distribution of households** by income is log-normal with normal parameters mu# and sigma#,

the **distribution of total income** by household income has a log-normal distribution where mu$ = mu# + sigma#^2; sigma$ = sigma#.

See Aitchison and Brown (1957), p. 158.
Special thanks to Mohammod Irfan (Denver University) for his help on this topic.

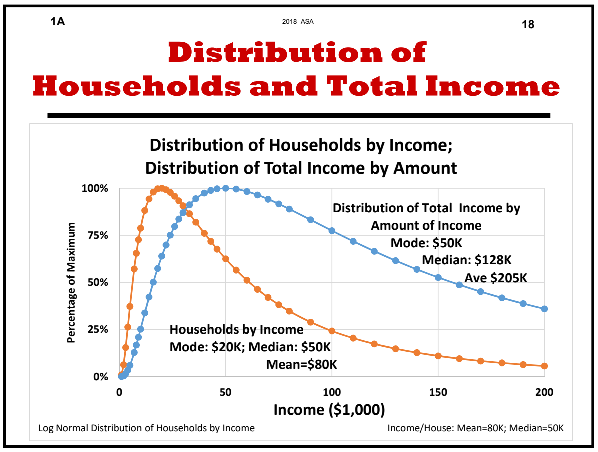


1A 2018 ASA 17

Distribution of Total Income

Distribution of Total Income by Income per Household

LogNormal Dist of Units by Income Median=50K; Mean=80K

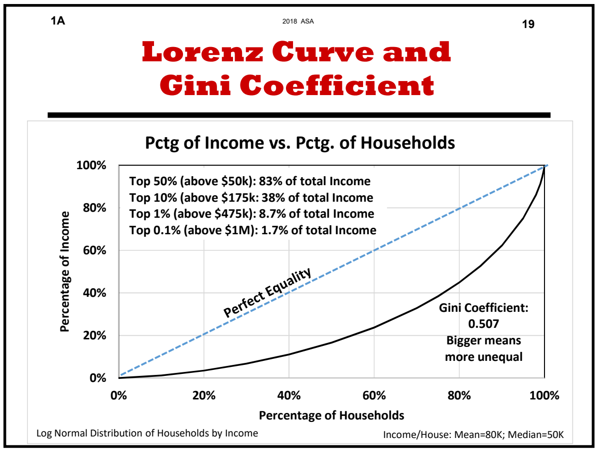


1A 2018 ASA 18

Distribution of Households and Total Income

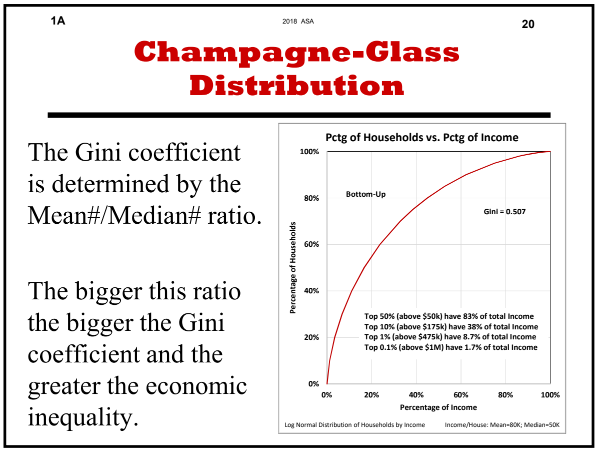
Distribution of Households by Income; Distribution of Total Income by Amount

Log Normal Distribution of Households by Income Income/House: Mean=80K; Median=50K



1A 2018 ASA 19

Lorenz Curve and Gini Coefficient

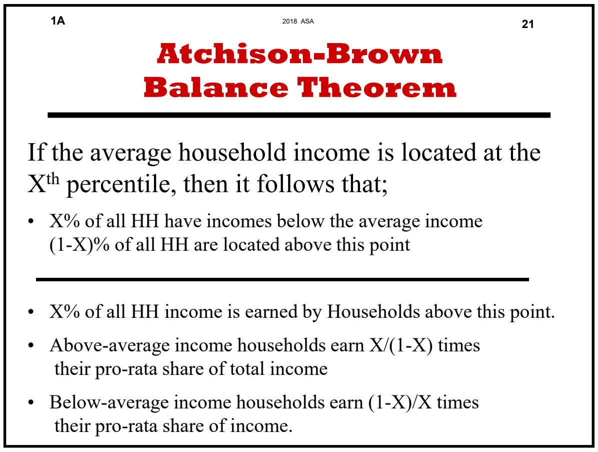


1A 2018 ASA 20

Champagne-Glass Distribution

The Gini coefficient is determined by the Mean#/Median# ratio.

The bigger this ratio the bigger the Gini coefficient and the greater the economic inequality.



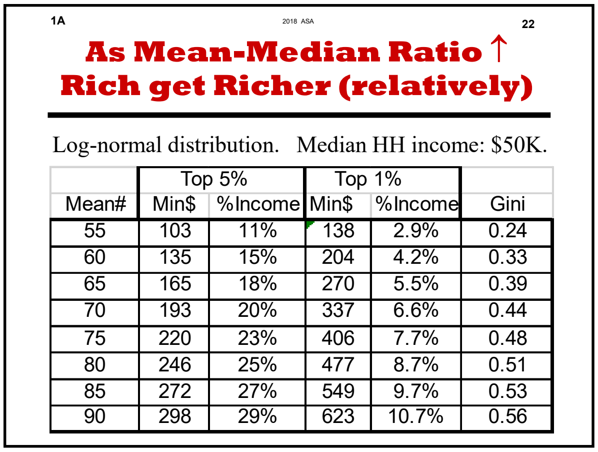
1A 2018 ASA 21

Atchison-Brown Balance Theorem

If the average household income is located at the Xth percentile, then it follows that;

- X% of all HH have incomes below the average income
  (1-X)% of all HH are located above this point

- X% of all HH income is earned by Households above this point.
- Above-average income households earn X/(1-X) times their pro-rata share of total income
- Below-average income households earn (1-X)/X times their pro-rata share of income.

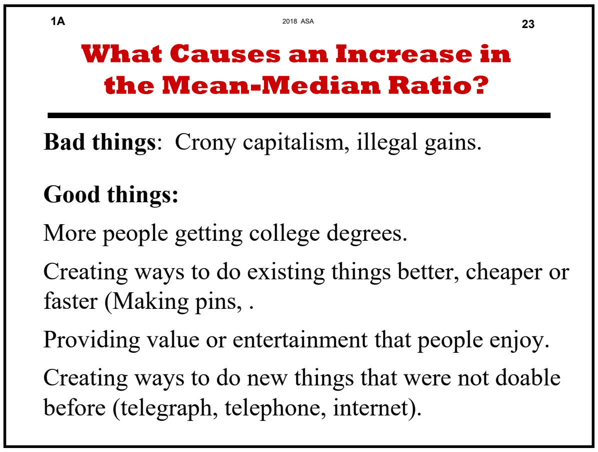


1A 2018 ASA 22

As Mean-Median Ratio ↑ Rich get Richer (relatively)

Log-normal distribution. Median HH income: $50K.

	Top 5%		Top 1%		
Mean#	Min$	%Income	Min$	%Income	Gini
55	103	11%	138	2.9%	0.24
60	135	15%	204	4.2%	0.33
65	165	18%	270	5.5%	0.39
70	193	20%	337	6.6%	0.44
75	220	23%	406	7.7%	0.48
80	246	25%	477	8.7%	0.51
85	272	27%	549	9.7%	0.53
90	298	29%	623	10.7%	0.56



1A 2018 ASA 23

What Causes an Increase in the Mean-Median Ratio?

Bad things: Crony capitalism, illegal gains.

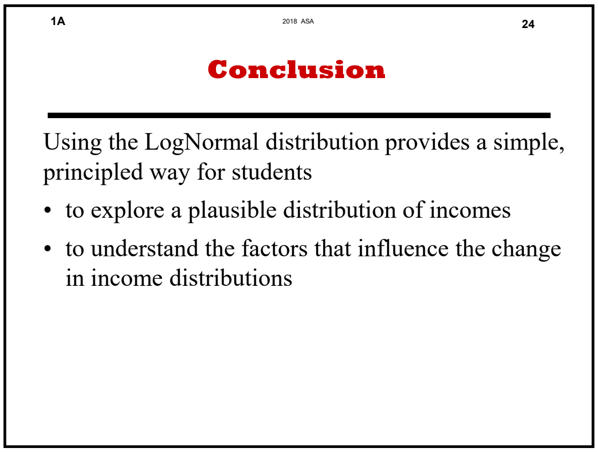
Good things:

More people getting college degrees.

Creating ways to do existing things better, cheaper or faster (Making pins, .

Providing value or entertainment that people enjoy.

Creating ways to do new things that were not doable before (telegraph, telephone, internet).



1A 2018 ASA 24

Conclusion

Using the LogNormal distribution provides a simple, principled way for students

- to explore a plausible distribution of incomes
- to understand the factors that influence the change in income distributions

Statistical Literacy: The Lognormal Distribution

Milo Schield, Augsburg U.

Editor: www.StatLit.org

US Rep: International Statistical Literacy Project

Amer. Statistical Association JSM

July 30, 2018

www.StatLit.org/

pdf/2018-Schield-ASA-Slides.pdf

XLS/Explore-LogNormal-Incomes-Excel2013.xlsx

Best-selling statistics books

80 million: *World Almanac (Since 1896)*

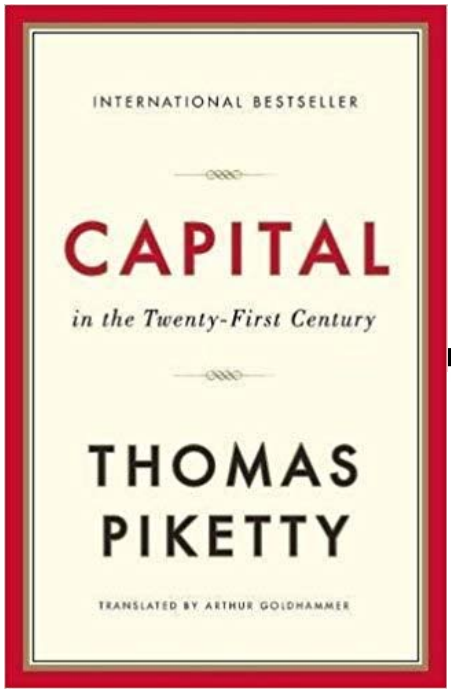
5 million: Economist: *World in Figures* (200K/yr; 25 years)

1.5 million: Piketty (2017): *Capital in the 21st Century*

500,000: Murray & Hernstein (1994): *The Bell Curve*

200,000: Hacker (1992): *"Two Nations: Black and White, Separate, Hostile, Unequal.“*

https://www.washingtonpost.com/archive/lifestyle/1995/09/22/black-and-white-read-all-over-the-hot-books-that-make-the-melting-pot-boil/ee1de9b5-a172-4dfd-bb7a-1eb1d6cf9d77/



INTERNATIONAL BESTSELLER

CAPITAL

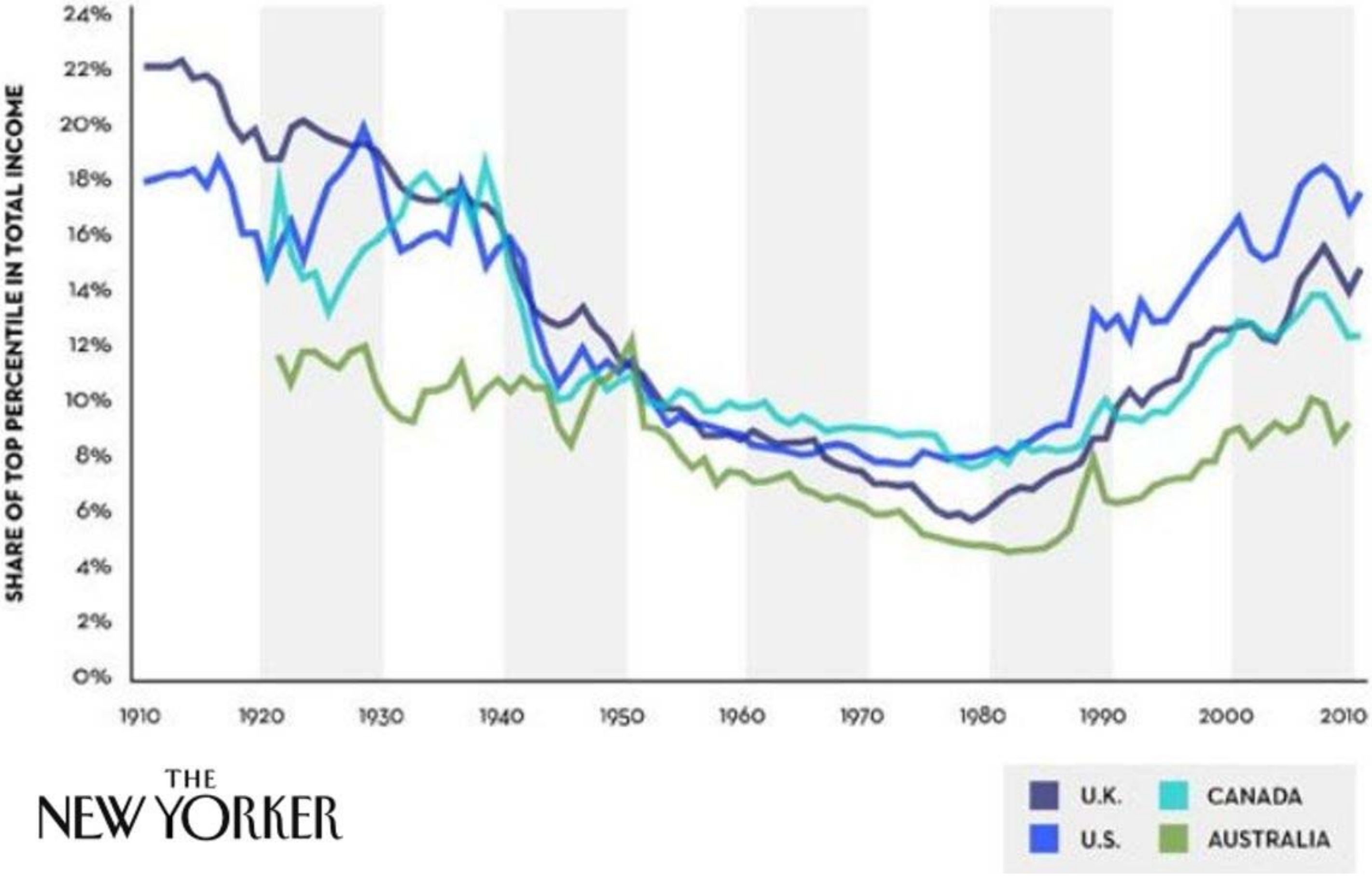
in the Twenty-First Century

THOMAS PIKETTY

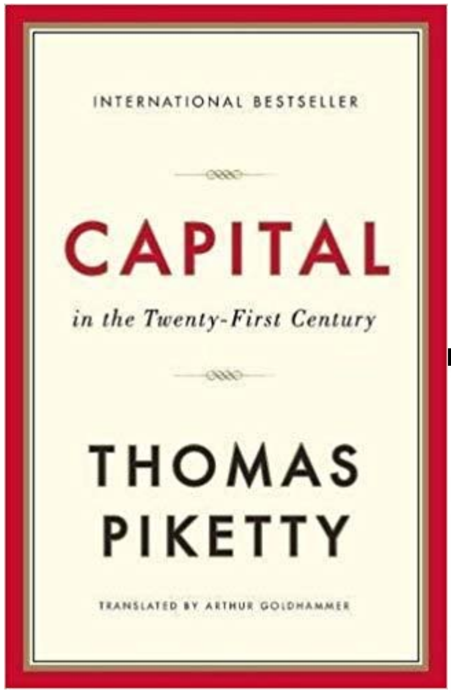
TRANSLATED BY ARTHUR GOLDHAMMER

Capital in the 21st Century: Income by Country (Top 1%)

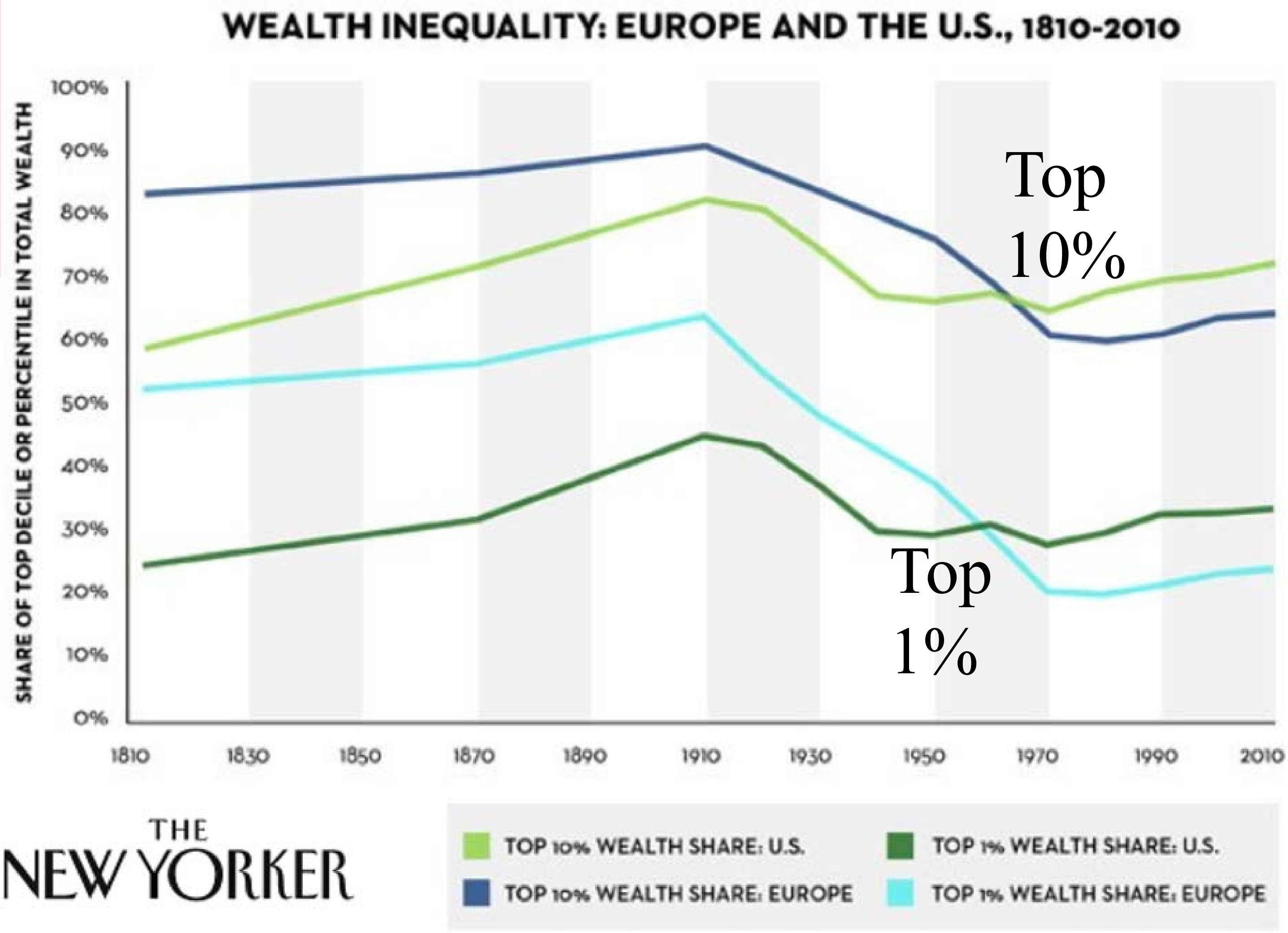
INCOME INEQUALITY IN ANGLO-SAXON COUNTRIES, 1910-2010



THE NEW YORKER



Capital in the 21st Century: Wealth by Country (Top 1%)

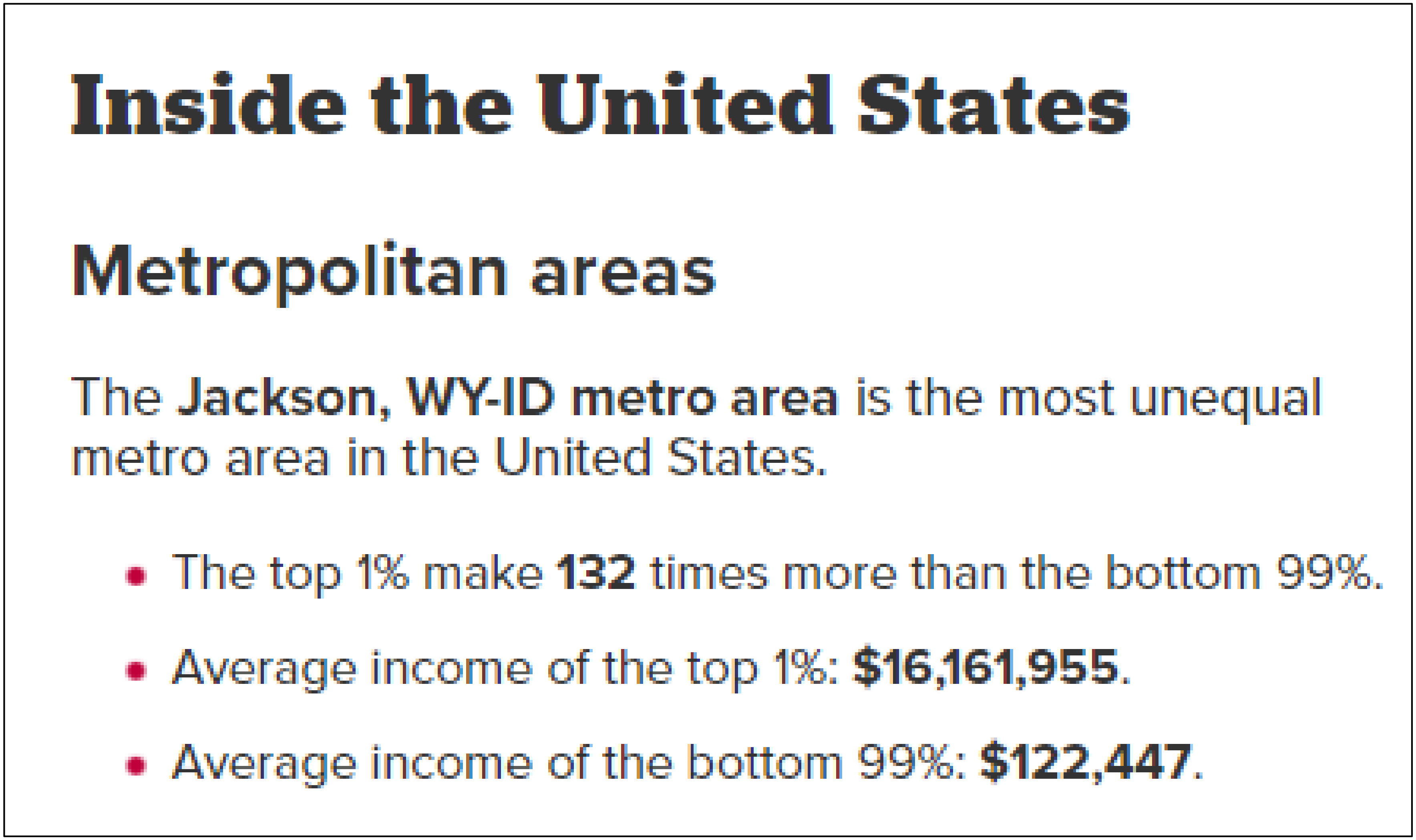


Top 10%

Top 1%

THE NEW YORKER

EPI (2018): US Income Inequality by Metro Area



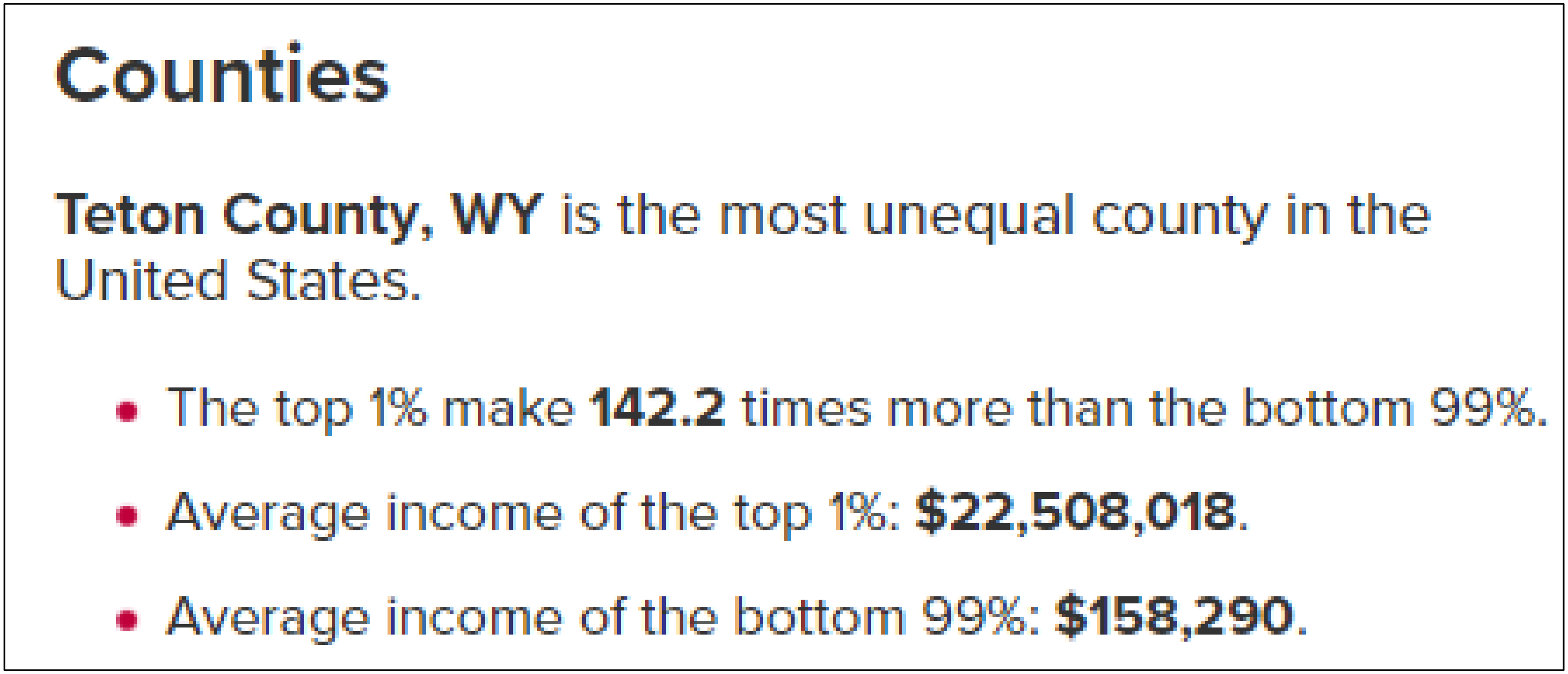
Inside the United States

Metropolitan areas

The **Jackson, WY-ID metro area** is the most unequal metro area in the United States.

- The top 1% make **132** times more than the bottom 99%.
- Average income of the top 1%: **$16,161,955**.
- Average income of the bottom 99%: **$122,447**.

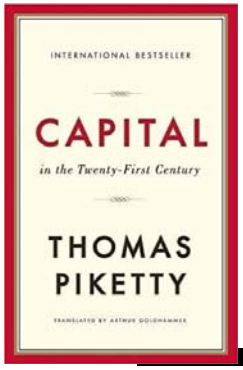
EPI (2018): US Income Inequality by County



Counties

Teton County, WY is the most unequal county in the United States.

- The top 1% make **142.2** times more than the bottom 99%.
- Average income of the top 1%: **$22,508,018**.
- Average income of the bottom 99%: **$158,290**.



Piketty: Censored Data Problem

Evaluate income share held by top 1% over time.

Data source: Tax data

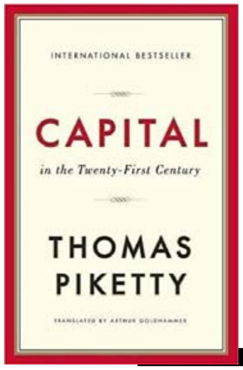
Problem: Tax authorities censors high-income data.

So, how did Piketty deduce the income share of top 1%

Piketty used a model: the Pareto distribution.

By fitting this model to uncensored incomes, he inferred the distribution of the censored incomes.

Atkinson et al (2011). P 12-14.



Piketty: *Capital in the 21st Century*

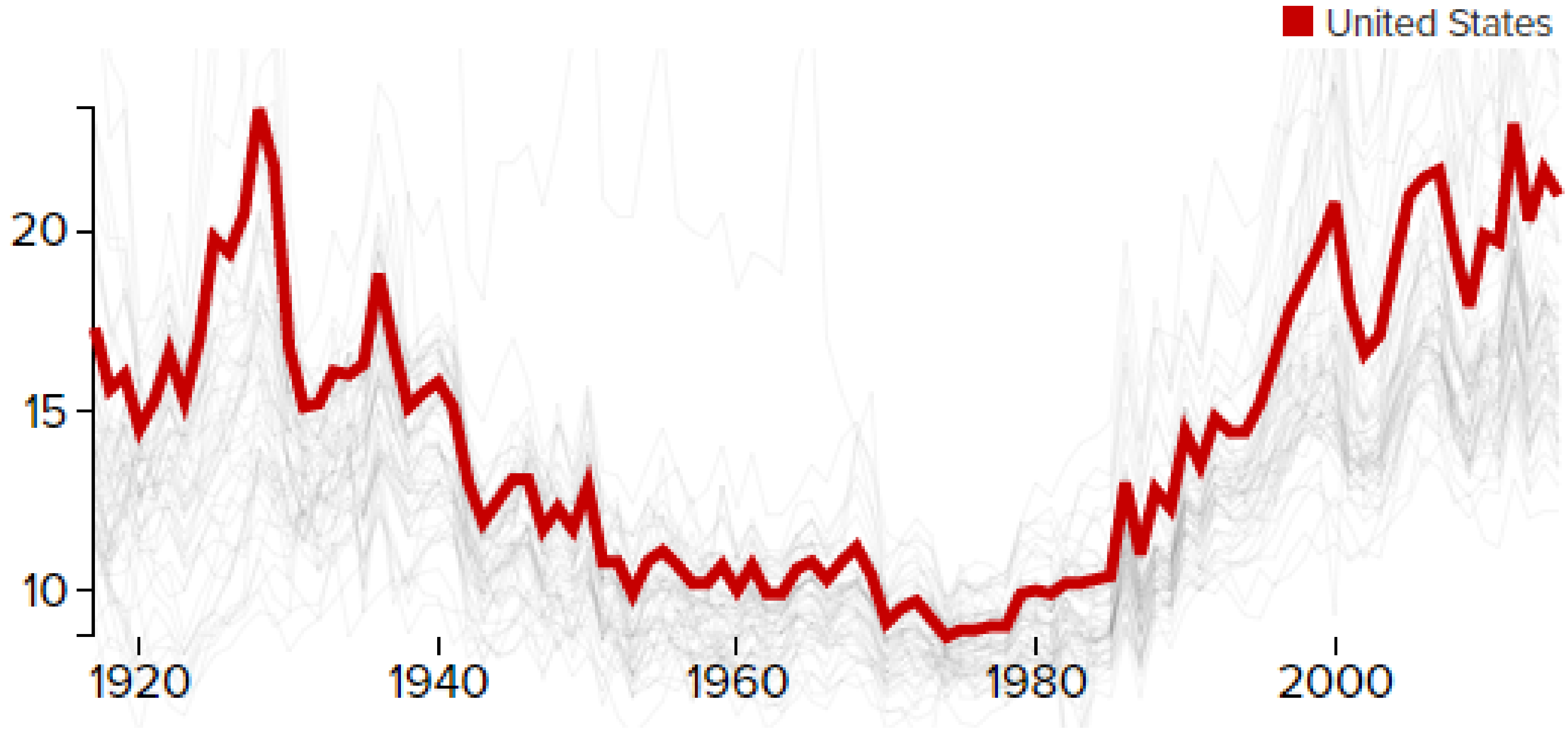
The key property of Pareto distributions is this: the “ratio of ‘average income y*(y) of individuals with income above y’ to y does not depend on the income threshold y.”

[Ave Income > y] / y = Beta

“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.”

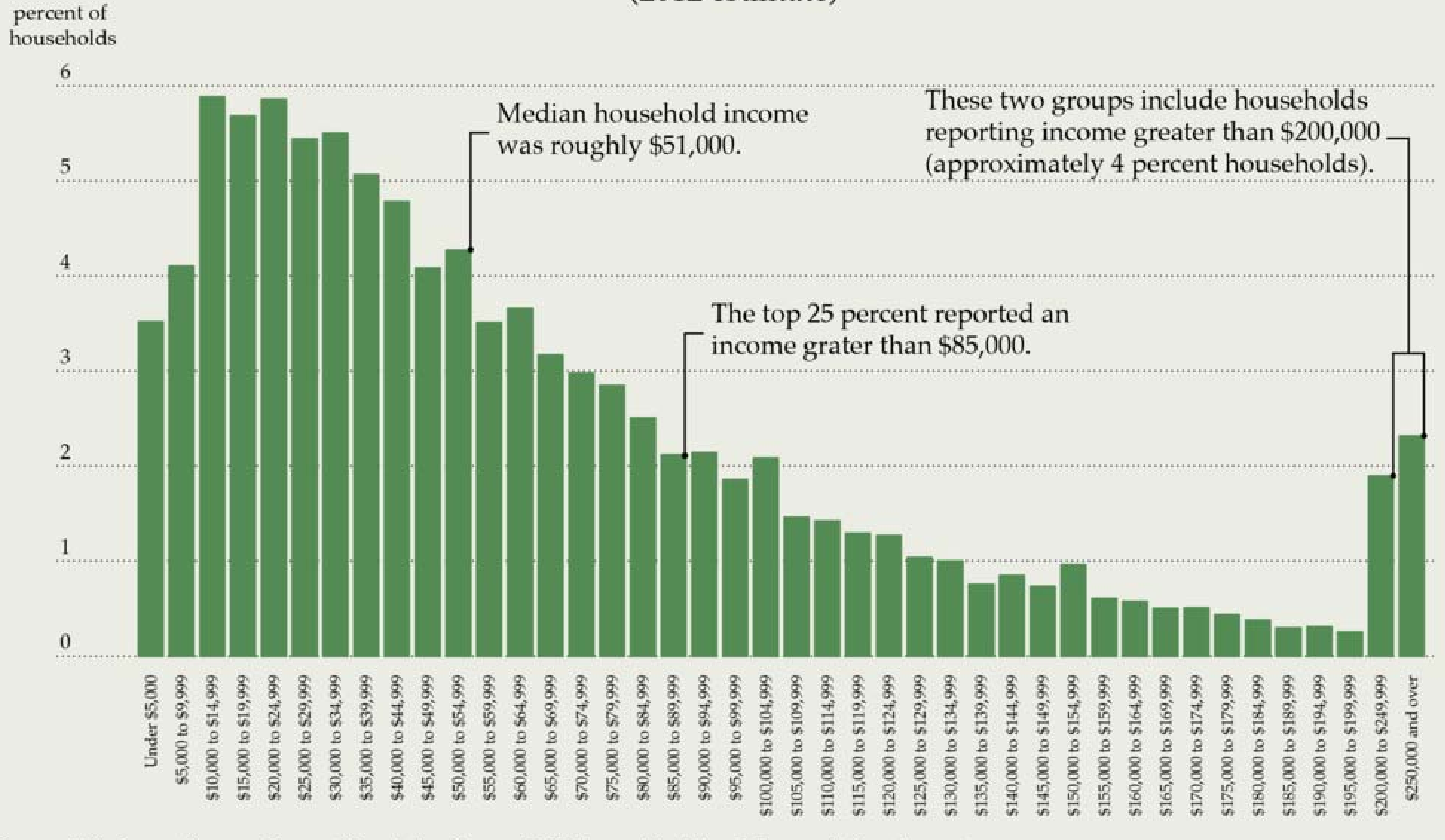
Atkinson, Piketty, Suan (2011). P 12-14.

EPI (2018): Income Inequality Top 1% Since 1920 by Country



Distribution of annual household income in the United States

(2012 estimate)



Median household income was roughly $51,000.

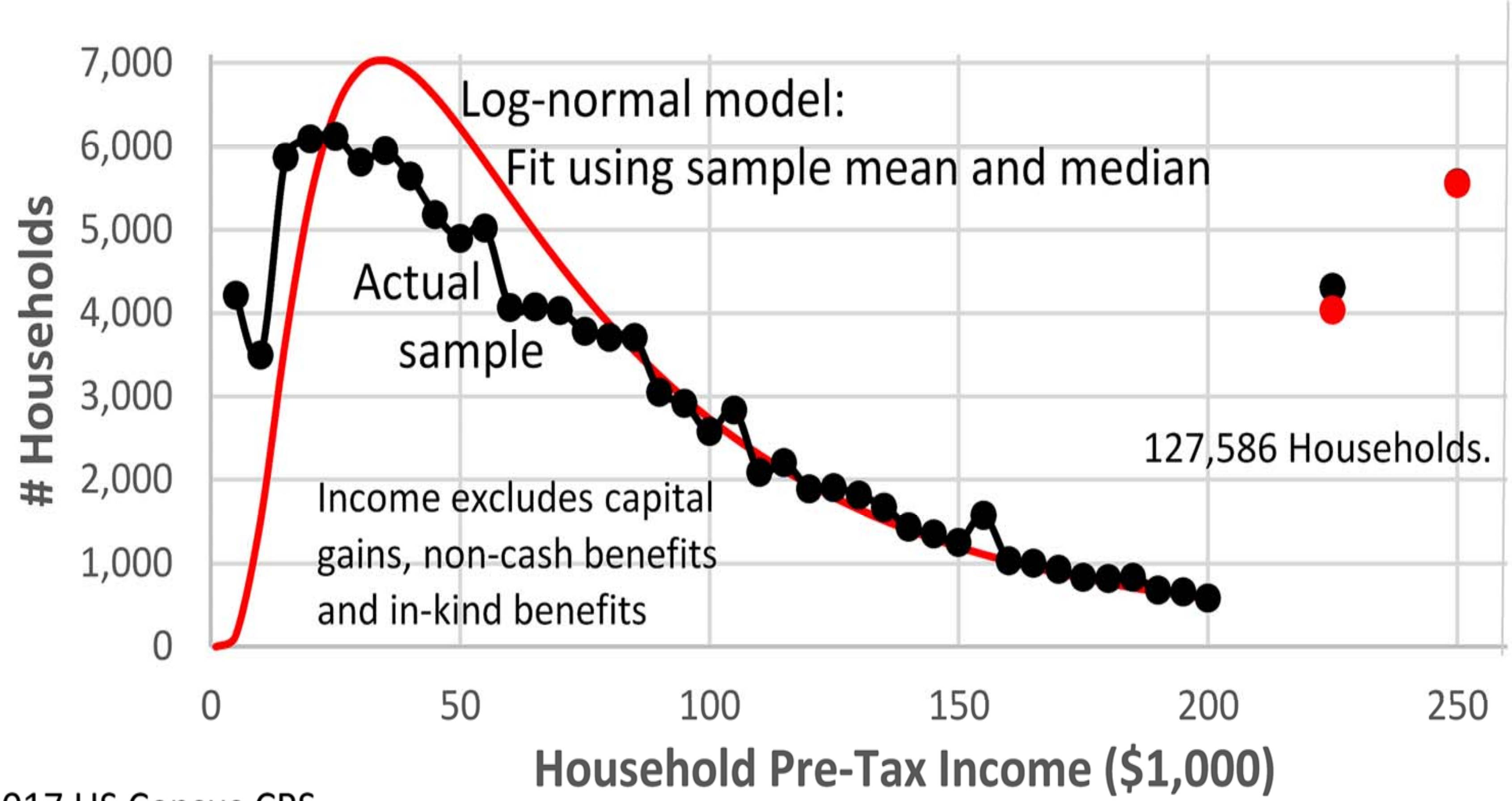
These two groups include households reporting income greater than $200,000 (approximately 4 percent households).

The top 25 percent reported an income grater than $85,000.

Source: U.S. Census Bureau, Current Population Survey, 2012 Annual Social and Economic Supplement

Distribution of Households by Income

US 2016: Median $61,710; Mean $86,220



Log-normal model:
Fit using sample mean and median

Actual
sample

127,586 Households.

Income excludes capital
gains, non-cash benefits
and in-kind benefits

2017 US Census CPS
www.census.gov/data/tables/time-series/demo/income-poverty/cps-hinc/hinc-06.2017.html

Log-Normal Distribution

Log-Normal shape is common. Examples:

- Incomes (bottom 97%), assets, size of cities
- Weight and blood pressure of humans (by gender)
- Stock and portfolio returns

Log-Normal is useful.

- Function is easier to work with than a histogram
- Understand what determines or explains shape
- calculate the share of total income held by the top X%
- calculate share of total income held by the ‘above-average’
- explore effects of change in mean-median ratio.

Log-Normal Distribution: Atchison and Brown

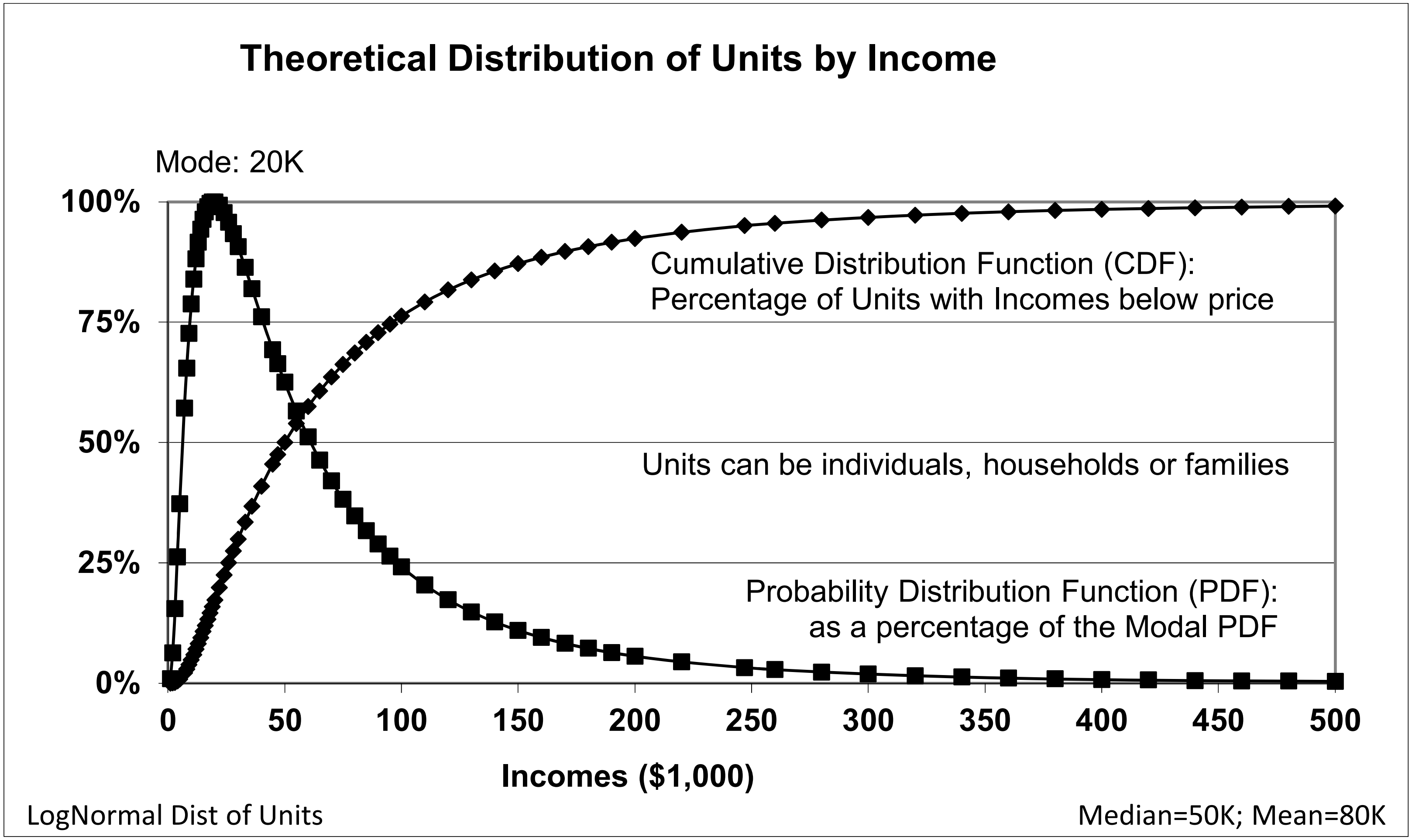
“In many ways, it [the Log-Normal] has remained the Cinderella of distributions, the interest of writers in the learned journals being curiously sporadic and that of the authors of statistical text-books but faintly aroused.”

“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.”

Shape is determined by the mean-median ratio.

Aitchison and Brown (1957). P 1.

Log-Normal Distribution of Units



Paired Distributions

For anything that is distributed by X, there are always two distributions:

1. Distribution of subjects by X
2. Distribution of total X by X.

Sometime we ignore the 2nd: height or weight.

Sometimes we care about the 2nd: income or assets.

Surprise: If the 1st is lognormal, so is the 2nd.

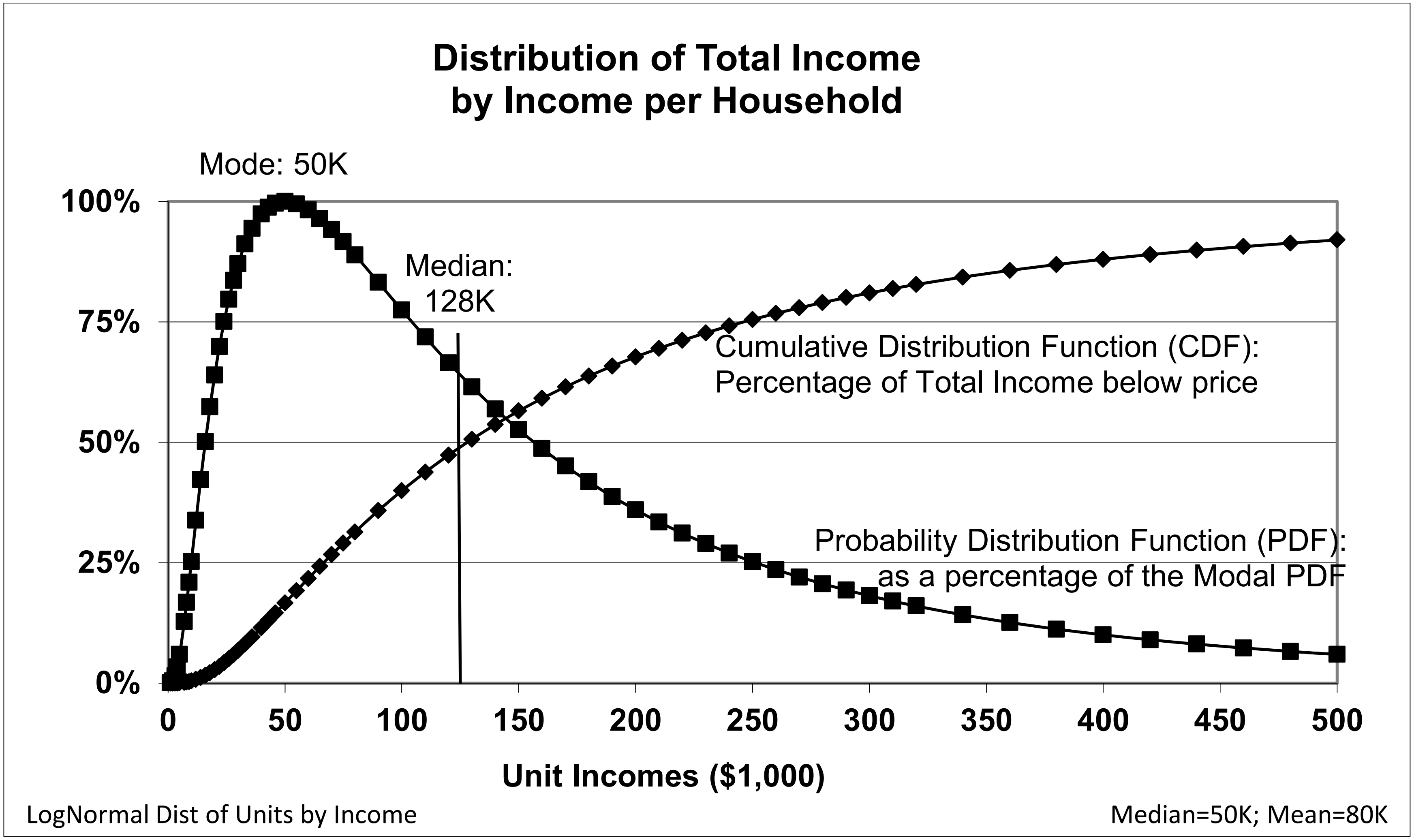
Distribution of Households and Total Income by Income

If the **distribution of households** by income is log-normal with normal parameters mu# and sigma#,

the **distribution of total income** by household income has a log-normal distribution where
mu$ = mu# + sigma#^2;   sigma$ = sigma#.

See Aitchison and Brown (1957), p. 158.
Special thanks to Mohammod Irfan (Denver University) for his help on this topic.

Distribution of Total Income

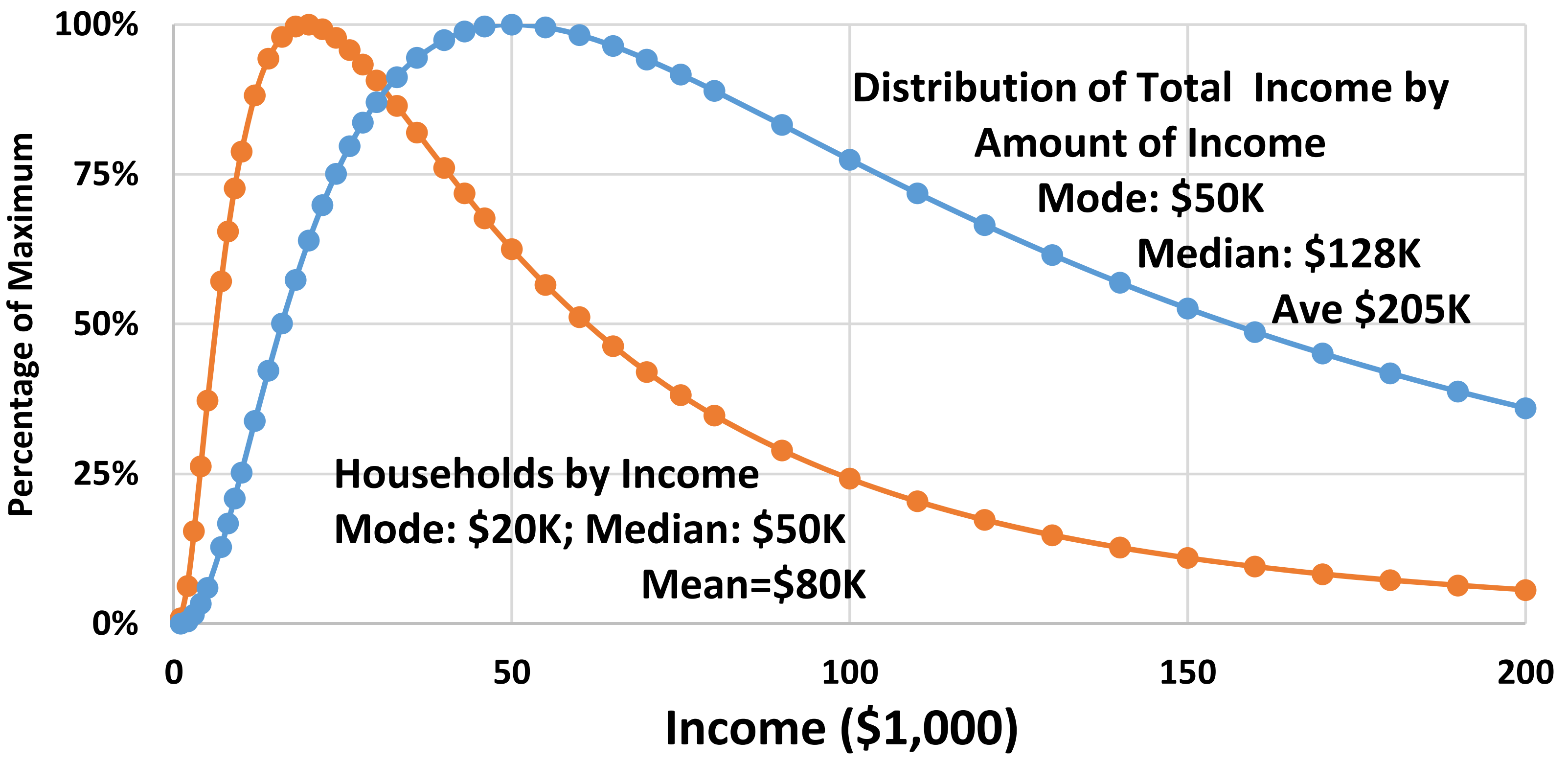


LogNormal Dist of Units by Income

Median=50K; Mean=80K

Distribution of Households and Total Income

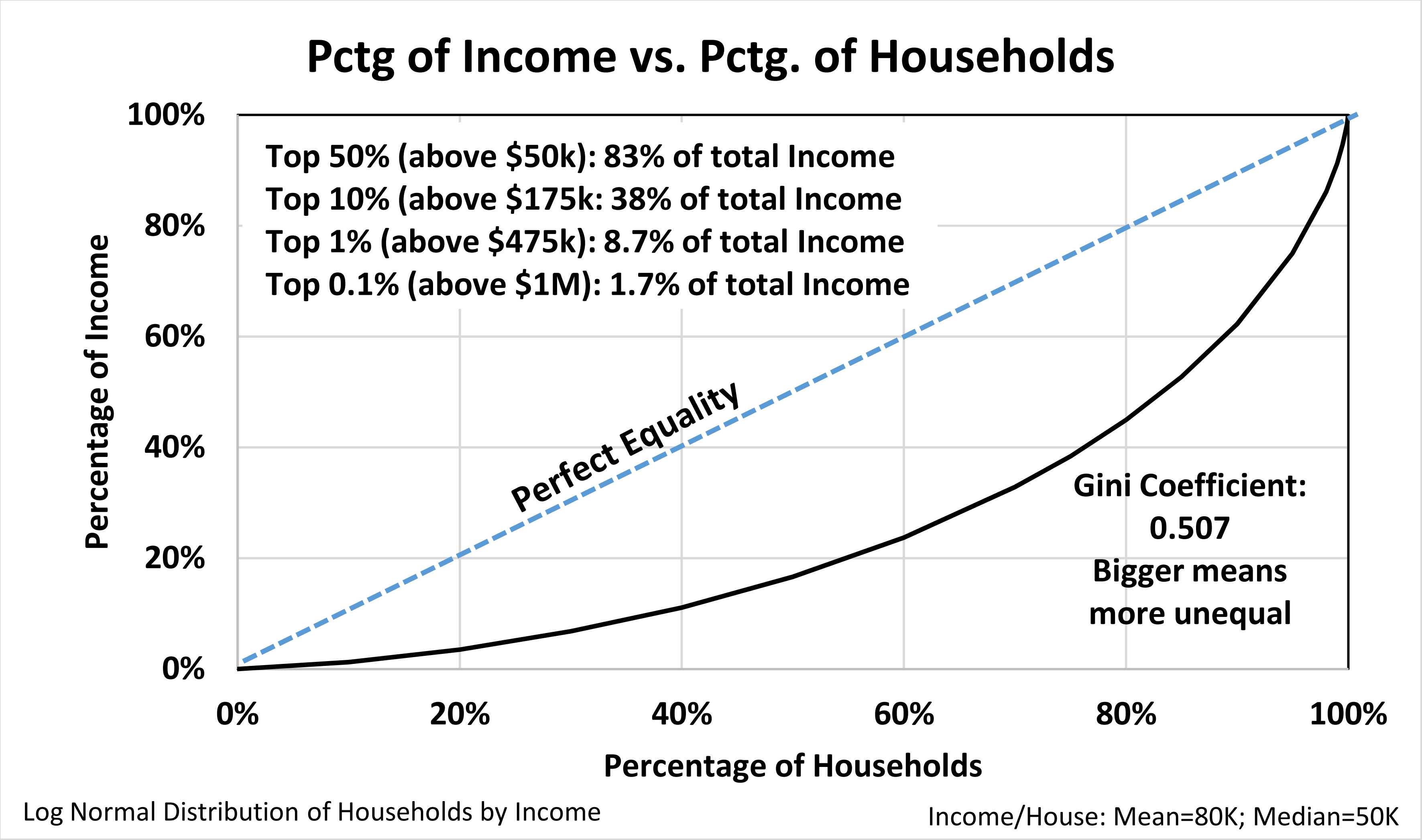
Distribution of Households by Income;
Distribution of Total Income by Amount



Log Normal Distribution of Households by Income

Income/House: Mean=80K; Median=50K

Lorenz Curve and Gini Coefficient



Top 50% (above $50k): 83% of total Income
Top 10% (above $175k: 38% of total Income
Top 1% (above $475k): 8.7% of total Income
Top 0.1% (above $1M): 1.7% of total Income

Gini Coefficient:
0.507
Bigger means
more unequal

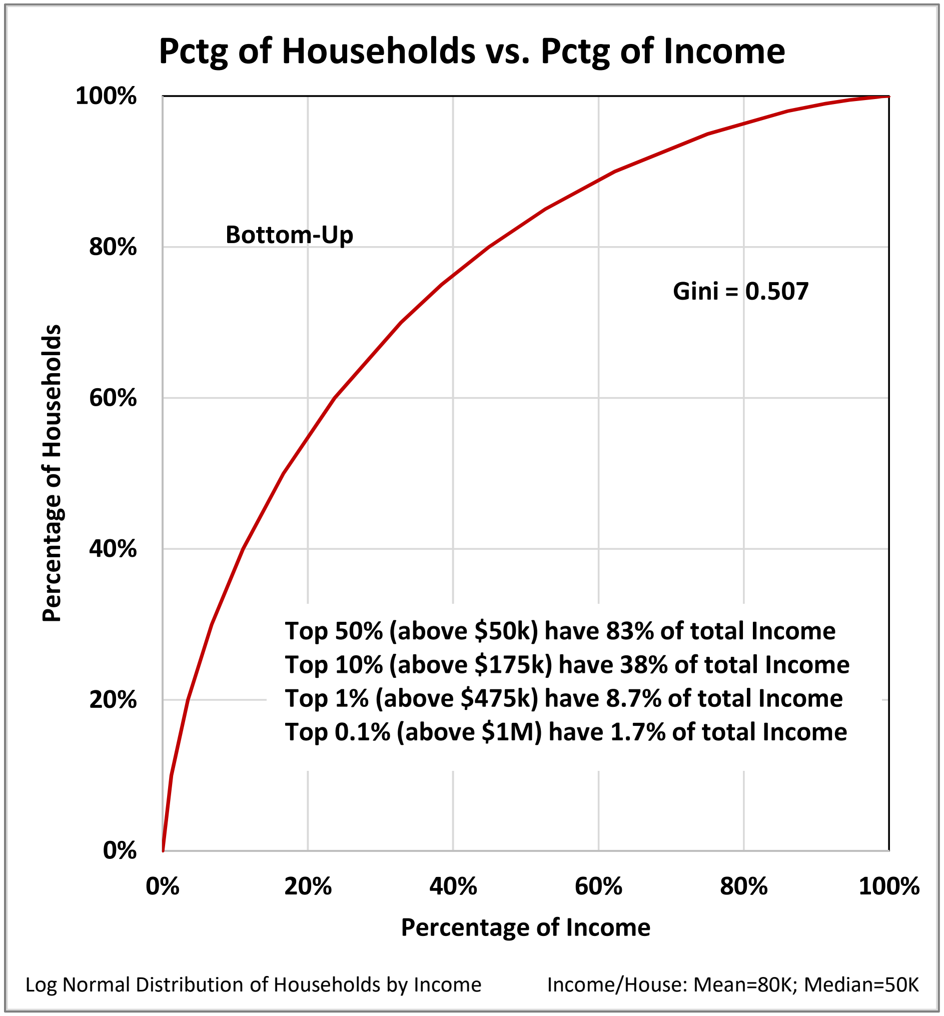
Log Normal Distribution of Households by Income

Income/House: Mean=80K; Median=50K

Champagne-Glass Distribution

The Gini coefficient is determined by the Mean#/Median# ratio.

The bigger this ratio the bigger the Gini coefficient and the greater the economic inequality.



Pctg of Households vs. Pctg of Income

Top 50% (above $50k) have 83% of total Income
Top 10% (above $175k) have 38% of total Income
Top 1% (above $475k) have 8.7% of total Income
Top 0.1% (above $1M) have 1.7% of total Income

Log Normal Distribution of Households by Income Income/House: Mean=80K; Median=50K

Atchison-Brown Balance Theorem

If the average household income is located at the Xth percentile, then it follows that;

- X% of all HH have incomes below the average income (1-X)% of all HH are located above this point

- X% of all HH income is earned by Households above this point.
- Above-average income households earn X/(1-X) times their pro-rata share of total income
- Below-average income households earn (1-X)/X times their pro-rata share of income.

As Mean-Median Ratio ↑ Rich get Richer (relatively)

Log-normal distribution. Median HH income: $50K.

	Top 5%		Top 1%		
Mean#	Min$	%Income	Min$	%Income	Gini
55	103	11%	138	2.9%	0.24
60	135	15%	204	4.2%	0.33
65	165	18%	270	5.5%	0.39
70	193	20%	337	6.6%	0.44
75	220	23%	406	7.7%	0.48
80	246	25%	477	8.7%	0.51
85	272	27%	549	9.7%	0.53
90	298	29%	623	10.7%	0.56

What Causes an Increase in the Mean-Median Ratio?

Bad things: Crony capitalism, illegal gains.

Good things:

More people getting college degrees.

Creating ways to do existing things better, cheaper or faster (Making pins, .

Providing value or entertainment that people enjoy.

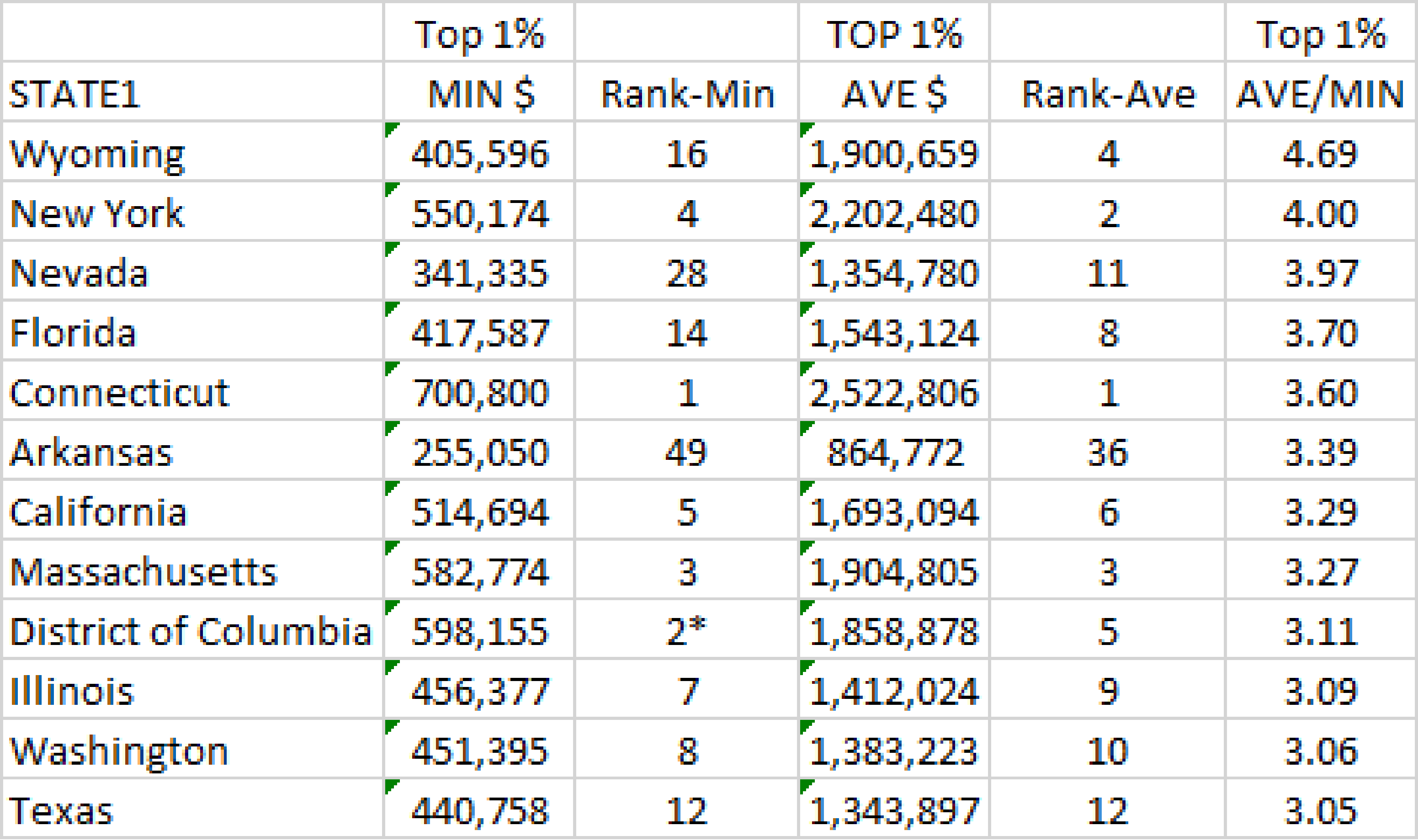
Creating ways to do new things that were not doable before (telegraph, telephone, internet).

Conclusion

Using the LogNormal distribution provides a simple, principled way for students

- to explore a plausible distribution of incomes
- to understand the factors that influence the change in income distributions

EPI (2018): US Income Inequality by State



	Top 1%		TOP 1%		Top 1%
STATE1	MIN $	Rank-Min	AVE $	Rank-Ave	AVE/MIN
Wyoming	405,596	16	1,900,659	4	4.69
New York	550,174	4	2,202,480	2	4.00
Nevada	341,335	28	1,354,780	11	3.97
Florida	417,587	14	1,543,124	8	3.70
Connecticut	700,800	1	2,522,806	1	3.60
Arkansas	255,050	49	864,772	36	3.39
California	514,694	5	1,693,094	6	3.29
Massachusetts	582,774	3	1,904,805	3	3.27
District of Columbia	598,155	2*	1,858,878	5	3.11
Illinois	456,377	7	1,412,024	9	3.09
Washington	451,395	8	1,383,223	10	3.06
Texas	440,758	12	1,343,897	12	3.05

Bibliography

Aitchison J and JAC Brown (1957). *The Log-normal Distribution*. Cambridge (UK): Cambridge University Press. Searchable copy at Google Books: http://books.google.com/books?id=Kus8AAAAIAAJ

Cassidy, John (2014). Piketty's Inequality Story in 6 Graphs. The New Yorker www.newyorker.com/news/john-cassidy/pikettys-inequality-story-in-six-charts

Cobham, Alex and Andy Sumner (2014). Is inequality all about the tails?: The Palma measure of income inequality. Significance. Volume 11 Issue 1.

Limpert, E., W.A. Stahel and M. Abbt (2001). Log-normal Distributions across the Sciences: Keys and Clues. *Bioscience* 51, No 5, May 2001, 342-352. Copy at http://stat.ethz.ch/~stahel/lognormal/bioscience.pdf

Schield, Milo (2013) Creating a Log-Normal Distribution using Excel 2013. www.statlit.org/pdf/Create-LogNormal-Excel2013-Demo-6up.pdf

Stahel, Werner (2014). Website: http://stat.ethz.ch/~stahel

Univ. Denver (2014). Using the LogNormal Distribution. Copy at http://www.du.edu/ifs/help/understand/economy/poverty/lognormal.html

Wikipedia. LogNormal Distribution.