Statistical Models and Disparity Measures of Annual Incomes in the US

Dr. Jeffrey Dawson - Professor, Dept. of Biostatistics, University of Iowa

Jessica Lee - New York University

Elaina McGovern - Boise State University

Juleika Torres - University of Puerto Rico- Mayagüez

Iowa Summer Institute in Biostatistics 2023

Outline

01

Importance of income disparity

02

Construction of data

03

Parametric models

04

Results after transformations

05

Conclusion

06

Limitations and future directions

Income Disparity

Definition	Unequal distribution of money earned among individuals or groups within the society.
Factors	Differences in education, job opportunities, skills, and social or economic policies.
Effect	Creates a cycle where the rich get richer, while the poor struggle to improve their financial situation.

Income Disparity

Issue	Over time, this income disparity can lead to a gap between the rich and the poor.
Policies	Addressing income disparity requires policies that promote equal opportunities, education access, and job training.
Expectation	By reducing income disparity, societies can attempt for a fairer and more equitable distribution of wealth.

Objectives

- Appropriately model the data set, disparity metrics, and factors that affect them given a variety of transformations.
- Understand and interpret the disparity in income.
- Propose potential solutions to reduce income disparities.

Data Collection

Individual Income Percentile	2022
1%	$0
2%	$0
3%	$3
4%	$1,000
5%	$2,040
6%	$3,500
7%	$4,950
8%	$6,000
9%	$7,200
10%	$8,801

https://dqydj.com/average-median-top-individual-income-percentiles/

YEAR		0.001%	0.01%	0.10%
2019		$60,658,598	$12,623,539	$2,458,432
2018		$68,934,261	$13,576,286	$2,514,209
2017		$63,430,119	$12,899,070	$2,374,937
2016		$53,052,900	$10,963,921	$2,124,117
2015		$59,380,503	$11,930,649	$2,220,264
2014		$56,981,718	$11,407,987	$2,136,762
2013		$45,097,112	$9,460,540	$1,860,848
2012		$62,068,187	$12,104,014	$2,161,175
2011		$41,965,258	$8,830,028	$1,717,675
2010		$45,039,369	$8,762,618	$1,634,386
2009		$34,381,494	$7,206,540	$1,469,393
2008		$49,546,782	$10,097,827	$1,867,652
2007		$62,955,875	$12,747,384	$2,251,017
2006		$54,665,360	$11,649,460	$2,124,625
2005		$50,796,495	$10,738,867	$1,938,175

https://en.wikipedia.org/wiki/Income_in_the_United_States

- Data percentiles taken from:
  - United States Census Bureau's Annual ASEC Survey
  - Internal Revenue Service Income Statistics

- 100,000 values representing all of USA
  - Taking in consideration the following quantiles of individual income:
    - 0.00001 to 0.99999

Estimating Income Quantiles

0 - 99th percentile

- Created 1000 equally spaced data points between each given percentile by recursively adding a fixed increment
  - Linear interpolation

97%	$232,000
98%	$280,100
99%	$401,622

https://dqydj.com/average-median-top-individual-income-percentiles/

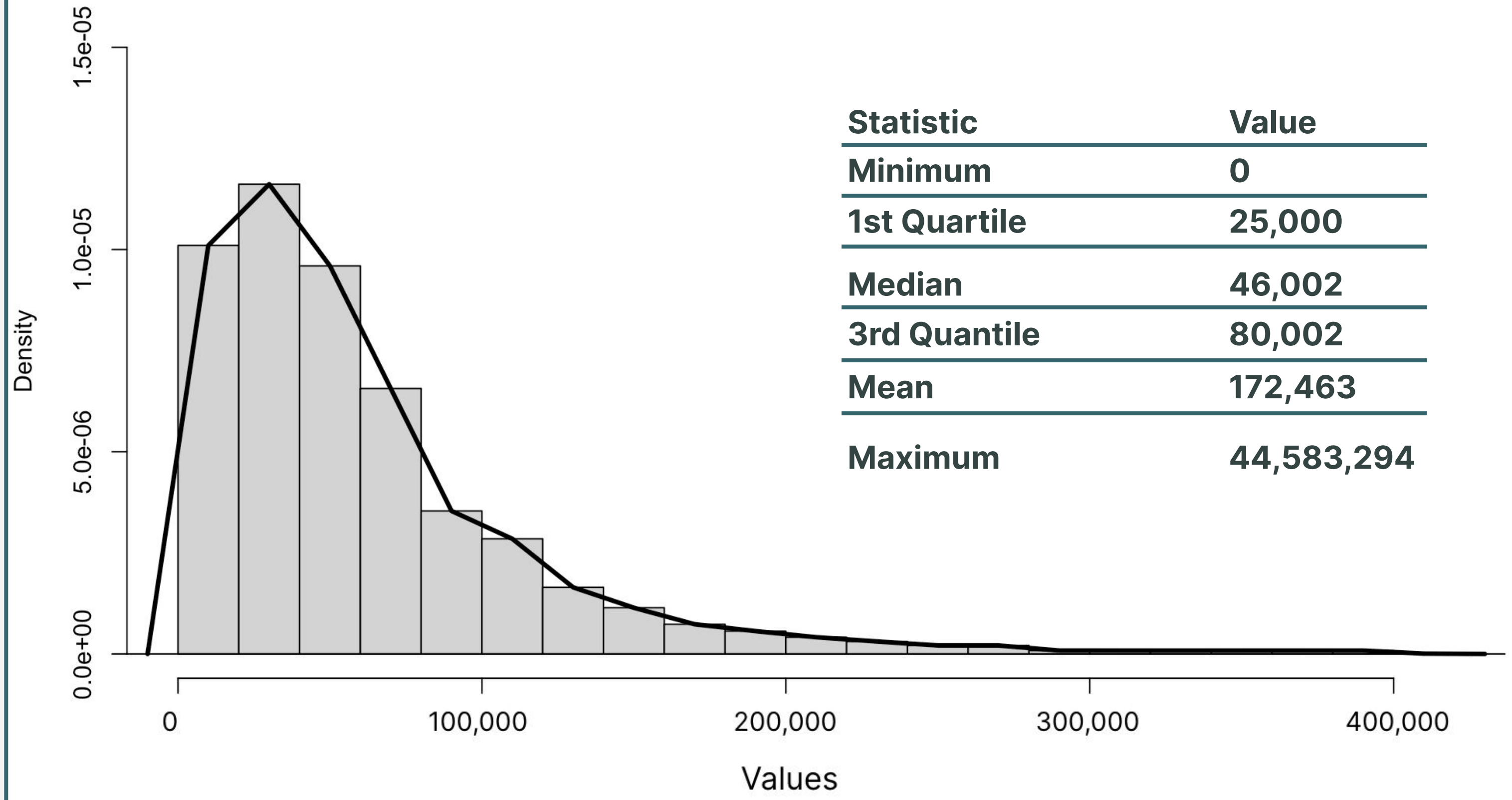
99 - 99.999th percentile

- Used curved projection to get reference points for individual income, then did linear interpolation
- Created 1000 total data points for the top 1%

YEAR		0.001%	0.01%	0.10%	1%
2019		$60,658,598	$12,623,539	$2,458,432	$546,434
2018		$68,934,261	$13,576,286	$2,514,209	$540,009
2017		$63,430,119	$12,899,070	$2,374,937	$515,371

https://en.wikipedia.org/wiki/Income_in_the_United_States

Histogram of Original Dataset



Statistic	Value
Minimum	0
1st Quartile	25,000
Median	46,002
3rd Quantile	80,002
Mean	172,463
Maximum	44,583,294

Measures of Disparity

Gini Index

A value between 0 and 1, where 0 is a perfectly equal society and 1 is perfectly unequal.

Decile Ratio

Value obtained from dividing the sum of the top 10% incomes by the sum of the bottom 10% of incomes.

Palma Ratio

Value obtained from dividing the sum of the top 10% incomes by the sum of the bottom 40% of incomes.

Coefficient of Variation

Value obtained from dividing the standard deviation of the income distribution by its mean.

Standard Deviation

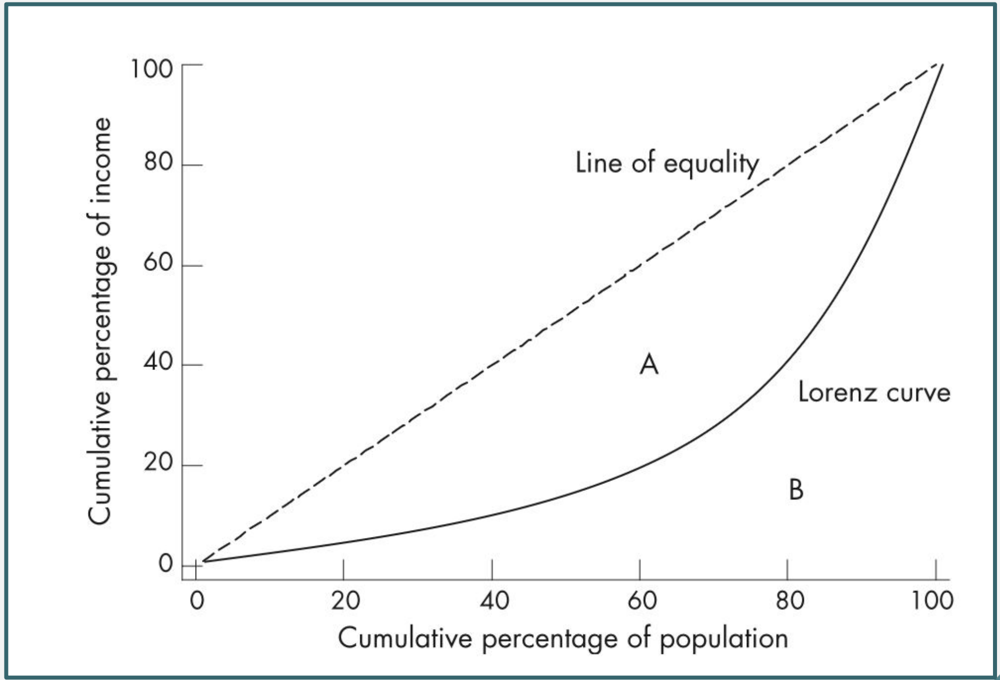
A measure of the amount of variability within a set of values.

Calculating the Gini Index

- The Lorenz curve shows the percentage of total income earned by cumulative percentage of the population.

- The Gini Index is the area between the line of equality and the Lorenz curve divided by the complete area under the line of equality.

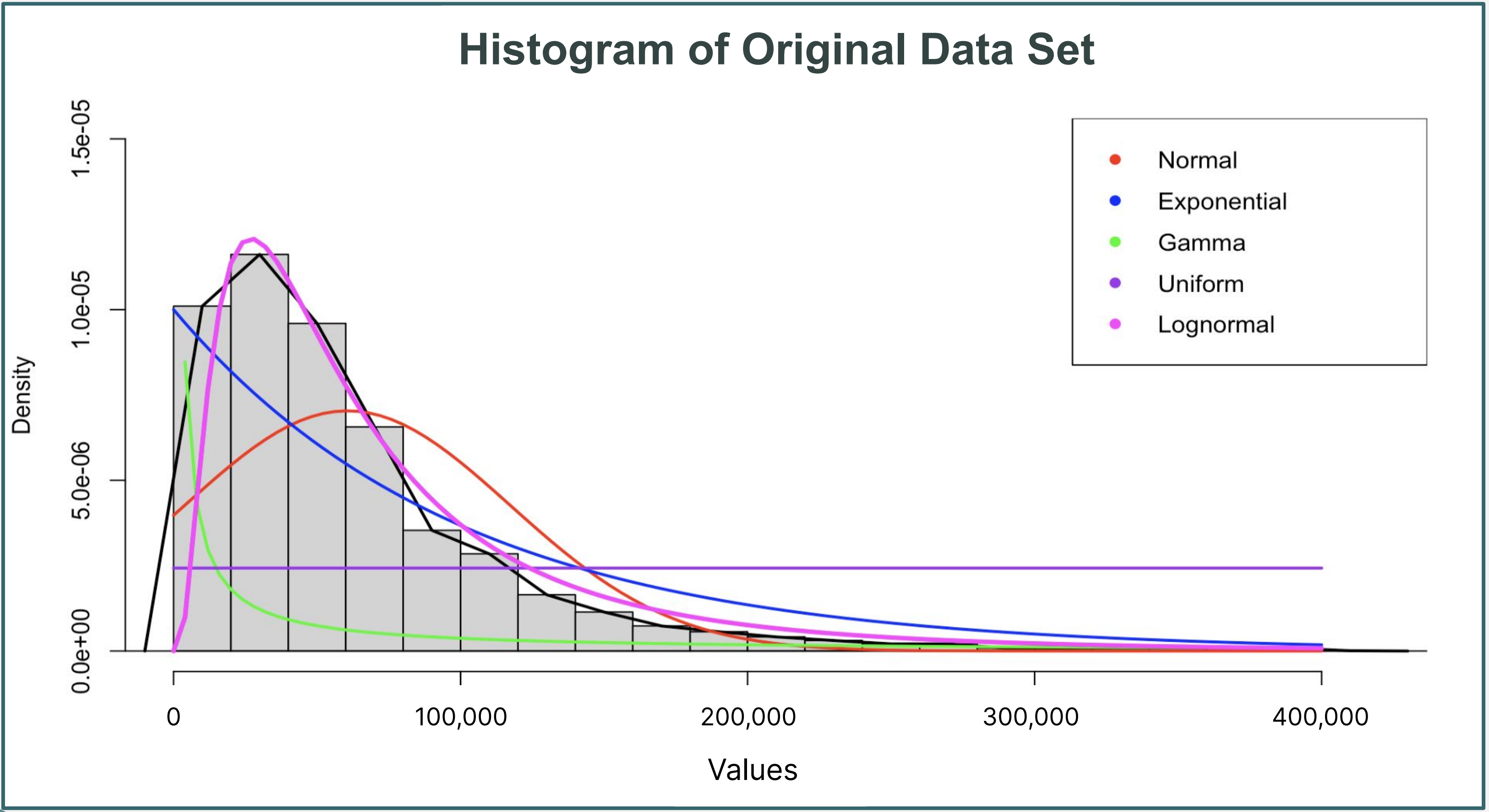
Gini coefficient = A / (A + B)



https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2652960/

Simulated Data Set Transformations

Dataset	Gini	SD	CV	Decile	Palma
Original	0.8038	1700098	9.8578	448.0569	17.4379
Add 5,000	0.7812	1700098	9.5800	165.4597	13.8104
Add 10,000	0.7598	1700098	9.3175	101.7640	11.4473
Add 20,000	0.7203	1700098	8.8334	57.7810	8.5533
Multiply 1.05	0.8038	1785103	9.8578	448.0569	17.4379
Multiply 1.10	0.8038	1870108	9.8578	448.0569	17.4379
Multiply 1.20	0.8038	2040118	9.8578	448.0569	17.4379
Minimum 5,000	0.8012	1700072	9.8428	236.7075	16.8523
Minimum 10,000	0.7967	1700028	9.8167	130.3751	15.9158
Minimum 20,000	0.7824	1699886	9.7302	65.1876	13.4193



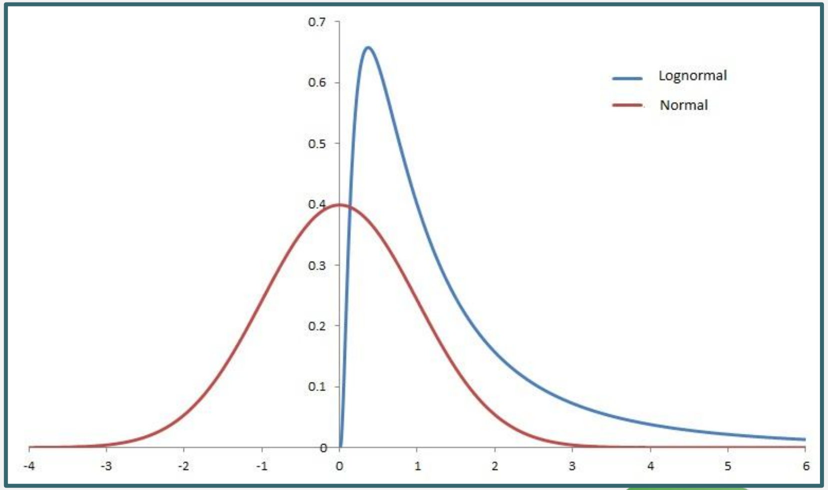
Histogram of Original Data Set
Normal
Exponential
Gamma
Uniform
Lognormal
Density
0.0e+00
5.0e-06
1.0e-05
1.5e-05
0
100,000
200,000
300,000
400,000
Values

Log-Normal Distribution

- A probability distribution where the values of a random variable, when transformed by taking their logarithms, exhibit a normal distribution.

- Distinguished by its right-skewed shape, featuring a longer tail on the right side of the distribution.

- Parameters:
  - The mean (μ)
    - Location parameter
  - Standard deviation (σ)
    - Scale parameter



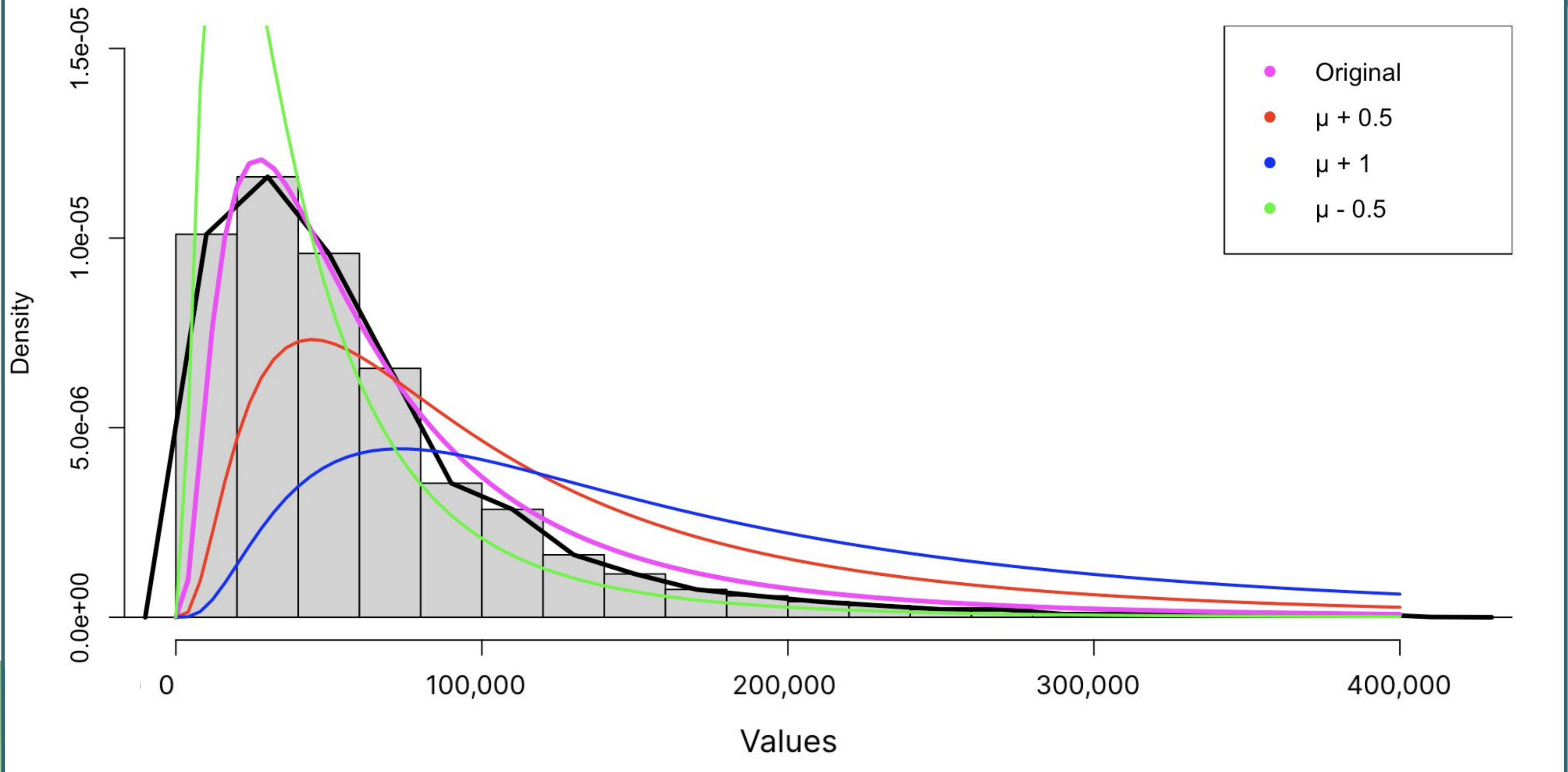
https://probabilityandstats.files.wordpress.com/2015/10/standard-normal-lognormal-pdfs.jpg

Adjusting to Log Scale

- Created a data set of all income values + 10,000 and took the natural log of all values in that data set
    - Added 10,000 to all income values to create a curve that matches the histogram more closely

- Calculated the μ and σ from the log-scale data set and took a random sample of 10,000 values that follow a log-normal distribution
    - $\mu = 10.9353$
    - $\sigma = 0.89$

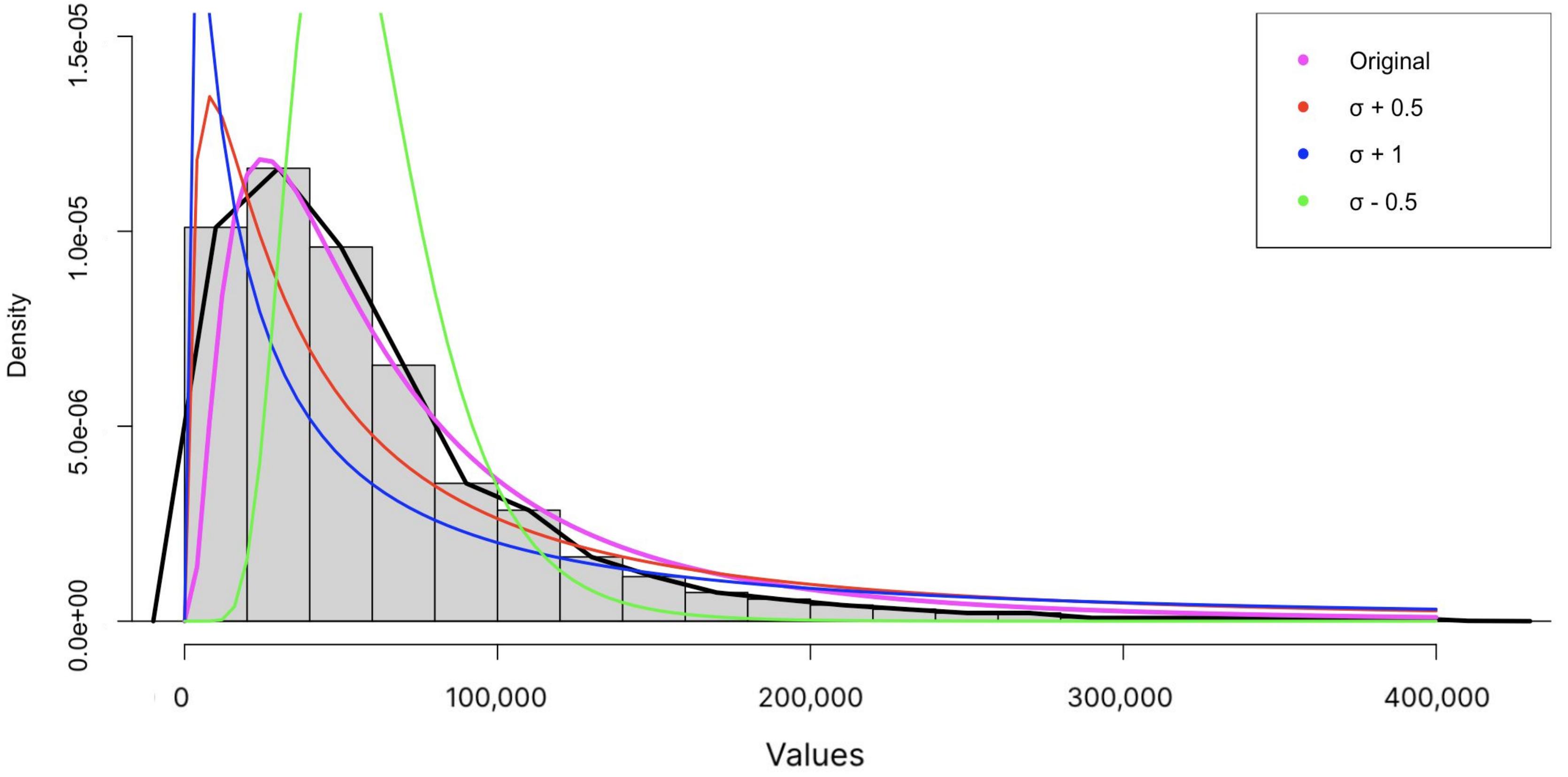
Histogram of Mu (μ) Addition Transformations



μ Addition Transformations: Log-Normal

Dataset	Gini	SD	CV	Decile	Palma
μ	0.4686	89477.93	1.0737	22.7103	2.7155
μ + 0.5	0.4686	147524.2	1.0737	22.7103	2.7155
μ + 1	0.4686	243226.2	1.0737	22.7103	2.7155
μ + 2	0.4686	661157.4	1.0737	22.7103	2.7155
μ - 0.5	0.4686	54271.11	1.0737	22.7103	2.7155
μ - 1	0.4686	32917.09	1.0737	22.7103	2.7155
μ - 2	0.4686	12109.52	1.0737	22.7103	2.7155

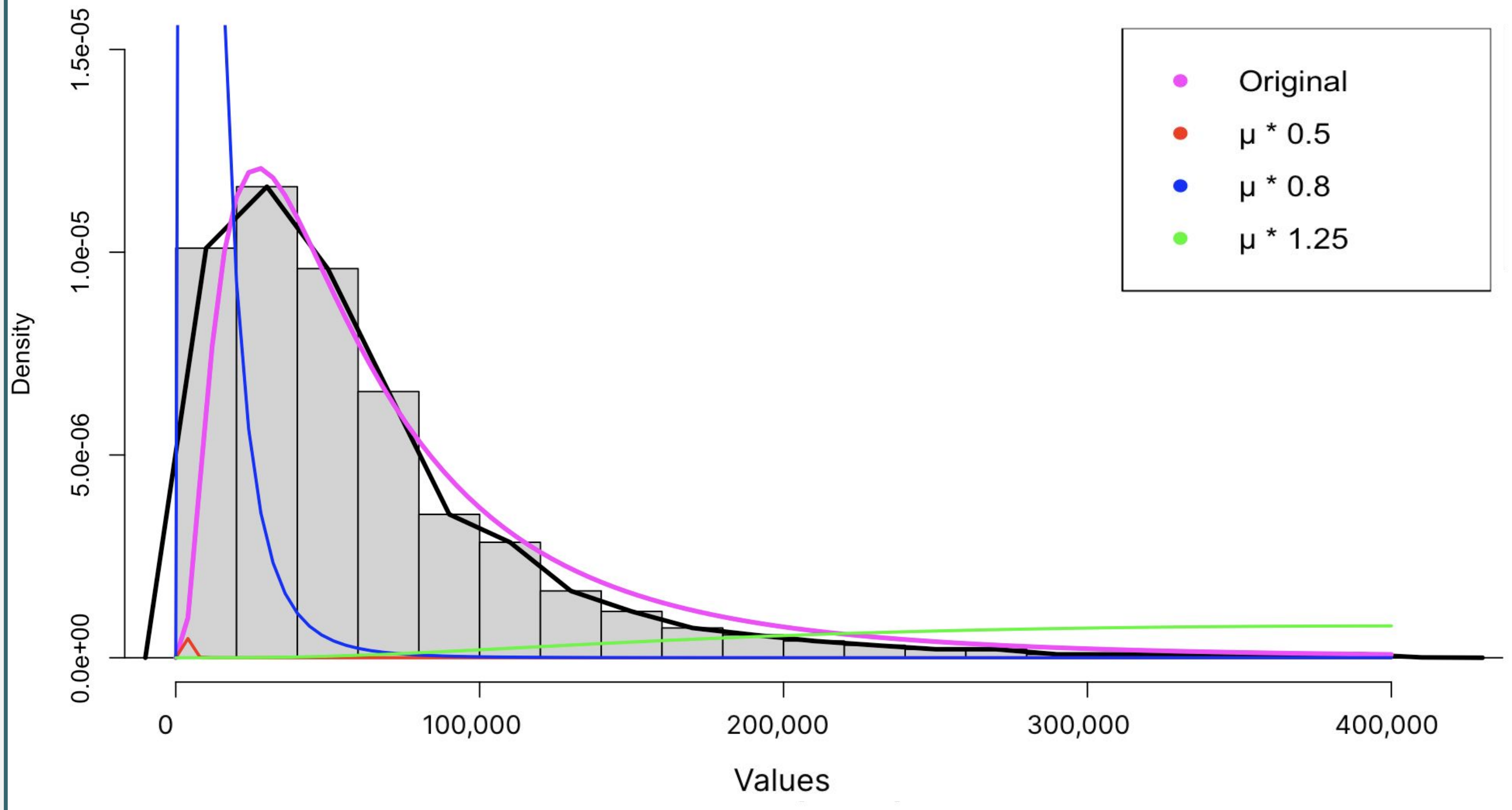
Histogram of Sigma (σ) Addition Transformations



σ Addition Transformations: Log-Normal

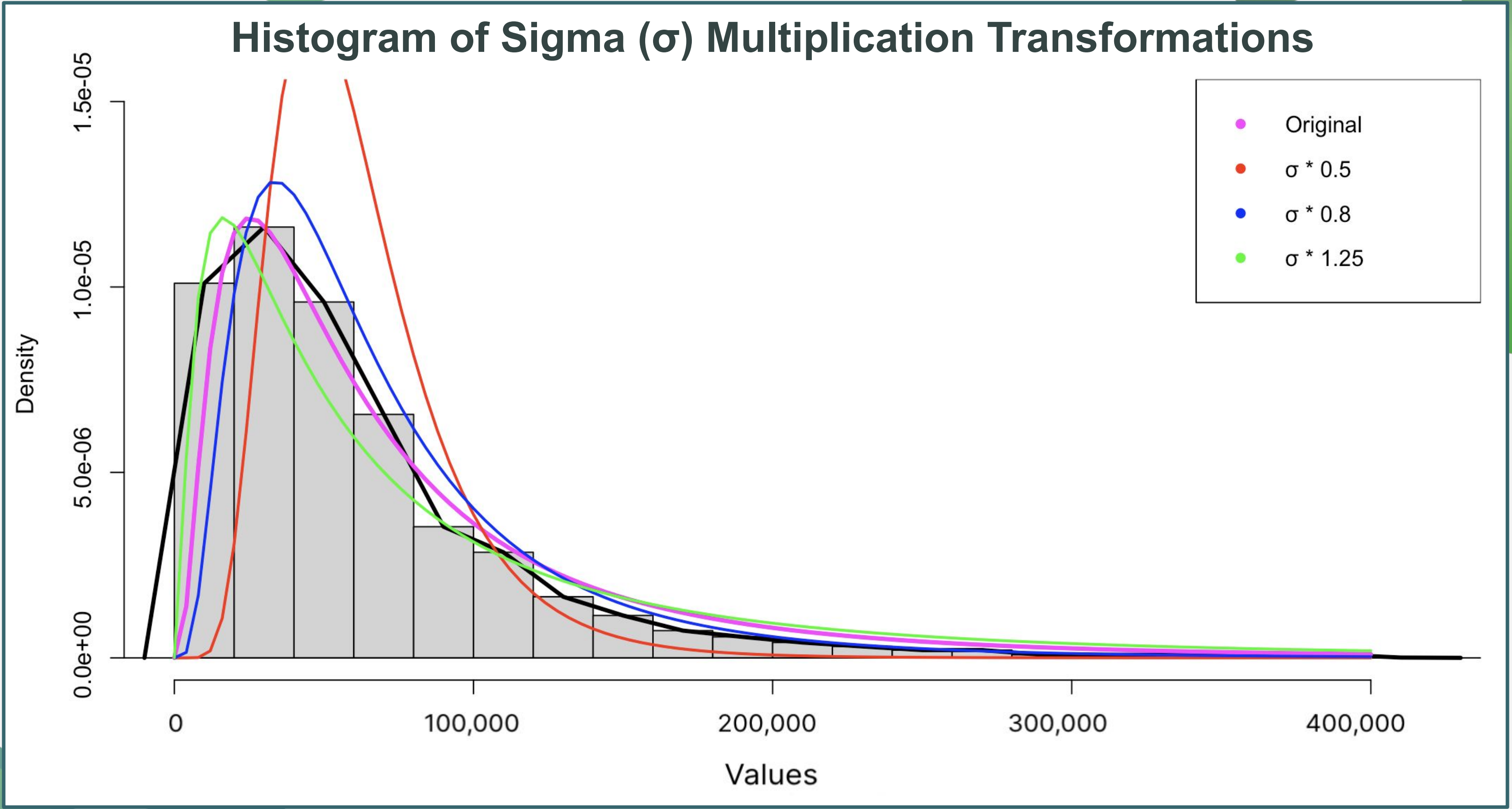
Dataset	Gini	SD	CV	Decile	Palma
σ	0.4686	89477.93	1.0737	22.7103	2.7155
σ + 0.5	0.6697	322039.9	2.2074	135.9916	10.5138
σ + 1	0.8111	1418688	4.4040	861.9352	42.6527
σ + 2	0.9493	47473036	16.1951	44755.03	888.8293
σ - 0.5	0.2169	24498.43	0.4040	3.9244	0.7153

Histogram of Mu (μ) Multiplication Transformations



μ Multiplication Transformations: Log-Normal

Dataset	Gini	SD	CV	Decile	Palma
μ	0.4686	89477.93	1.0737	22.7103	2.7155
μ * 0.1	0.4686	4.7587	1.0737	22.7103	2.7155
μ * 0.25	0.4686	24.5393	1.0737	22.7103	2.7155
μ * 0.5	0.4686	377.6996	1.0737	22.7103	2.7155
μ * 0.8	0.4686	10043.57	1.0737	22.7103	2.7155
μ * 1.25	0.4686	1377211	1.0737	22.7103	2.7155
μ * 1.5	0.4686	21197532	1.0737	22.7103	2.7155


Histogram of Sigma (σ) Multiplication Transformations
Original
σ * 0.5
σ * 0.8
σ * 1.25
Density
0.0e+00
5.0e-06
1.0e-05
1.5e-05
0
100,000
200,000
300,000
400,000
Values

σ Multiplication Transformations: Log-Normal

Dataset	Gini	SD	CV	Decile	Palma
σ	0.4686	89477.93	1.0737	22.7103	2.7155
σ * 0.1	0.0450	5007.523	0.0888	1.3649	0.3180
σ * 0.25	0.1245	12908.62	0.2242	2.1762	0.4554
σ * 0.5	0.2459	28820.55	0.4647	4.7393	0.8261
σ * 0.8	0.3836	57970.36	0.8014	12.1058	1.6858
σ * 1.25	0.5655	155366.8	1.4958	50.1727	4.9463
σ * 1.5	0.6509	278053.6	2.0480	111.8242	9.0673

Results

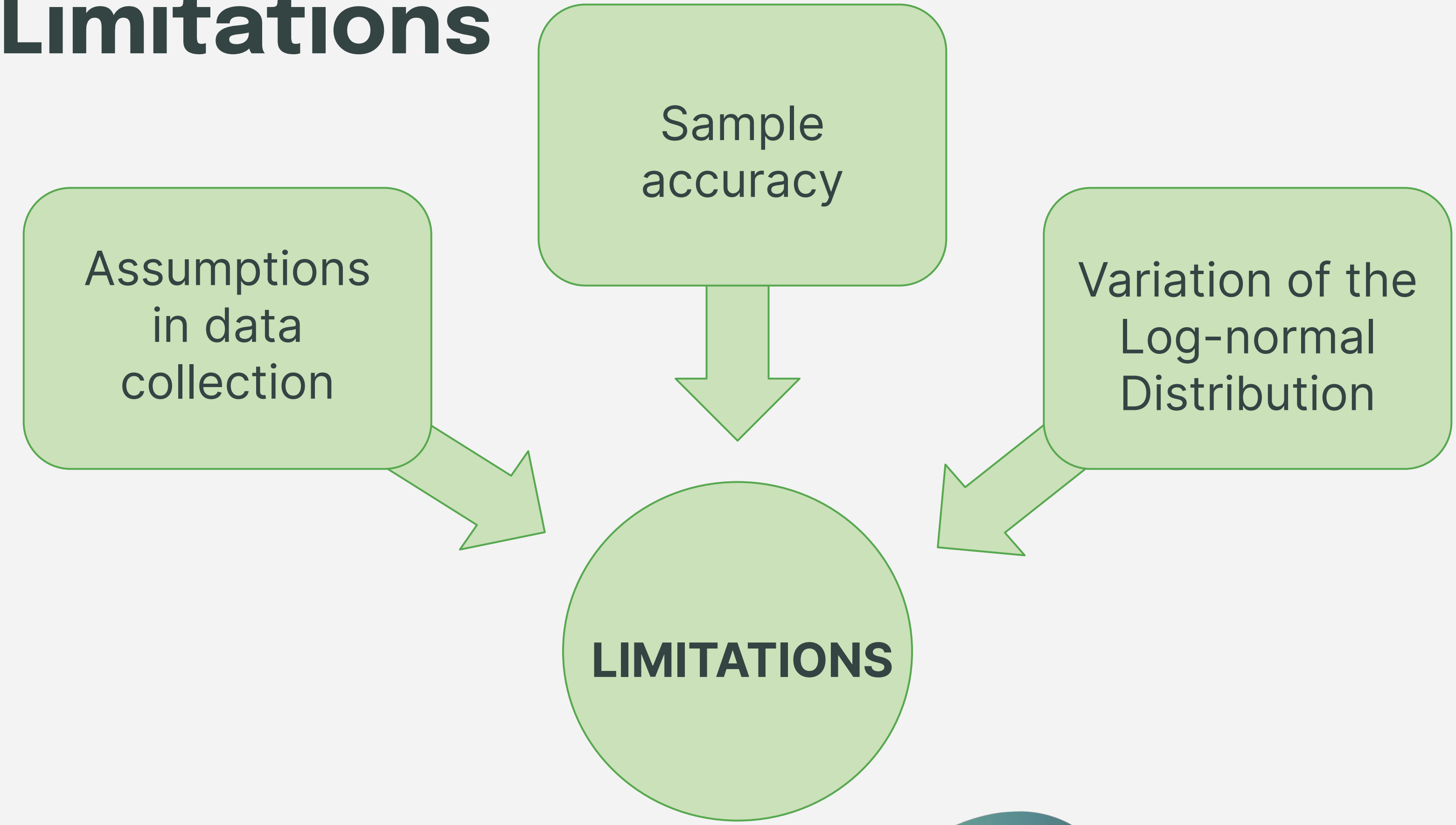
- Disparity measures, excluding the standard deviation, were unaffected by the additive and multiplicative changes to μ.
- Relatively small multiplicative changes to μ results in large variations of the original log-normal curve.
- Generally, multiplying numbers > 1 to the parameters will increase disparity measures while numbers < 1 will decrease them.
- Changing σ parameters on the log-normal distribution were more impactful in reducing disparities than changes to μ.

Conclusion

- Potential solutions to reduce disparity include redistributing wealth from the top earners to the bottom earners, or establishing a minimum yearly salary.
- It is important to consider levels of detail beyond modeled percentiles because of how much wealth is controlled by the top 1% while the bottom percentiles earn nothing.

https://www.salary.sg/2018/compare-your-household-income-2018/

Limitations



Future Directions

Other parametric models

Intersectionality with other factors

Longitudinal studies

Expansion of current work

Acknowledgements

The University of Iowa

http://surl.li/jflud

National Heart Lung and Blood Institute (NHLBI) Grant #HL161716-01

https://www.nhlbi.nih.gov/

- Jeffrey Dawson, ScD - Mentor
- Gideon Zamba, PhD - ISIB Director
- Terry Kirk - ISIB Coordinator
- Eliezer Santos - ISIB Teaching Assistant

References

(2022). *SOI Tax Stats - Individual statistical tables by tax rate and income percentile.* Internal Revenue Service. https://www.irs.gov/statistics/soi-tax-stats-individual-statistical-tables-by-tax-rate-and-income-percentile

(2022). *Average, Median, Top 1%, and all United States Individual Income Percentiles*. DQYDJ – Don't Quit Your Day Job. . . https://dqydj.com/average-median-top-individual-income-percentiles/

De Maio, F. (2007). Income inequality measures. *Journal of Epidemiology & Community Health*, *61*(10), 849-852.

Kawachi, I., & Subramanian, S. (2014). Income inequality. *Social epidemiology*, *126*, 126-152.

Siano, D. (1972). The log-normal distribution function. *Journal of Chemical Education*, *49*(11), 755.

Wilkinson, R., & Pickett, K. (2009). Income inequality and social dysfunction. *Annual review of sociology*, *35*, 493-511.

Thanks!

Gini Index Discrepancy

- Gini Index including top 1%: 0.8038
- Gini Index excluding top 1%: 0.4584
- Gini Index of log-normal data: 0.4686