6

Normal Isn't "Normal" when It Comes to Income

Ted Galanthay and Thomas J. Pfaff

In this two-day module, students confront some of the problems in calculating summary statistics of data that have been grouped into bins of varying widths. Through data exploration, students gain additional insights from attempting to calculate median and mean from summary data and visualizing heavily skewed data with a large range. The module also leads students to reconsider their perceptions of income inequality in the United States. In particular, students explore how skewed the U.S. income distribution really is. This module is appropriate for any statistics course or a course that incorporates statistical reasoning.

Keywords: descriptive statistics, income inequality, wealth inequality, log transformed data, histogram binning, histogram construction, discussion, think-pair-share

Students in introductory statistics courses learn basic properties of the normal or bell-shaped distribution while skewed distributions are only briefly, if ever, considered. As a result, when students think of how income is distributed, they often think in terms of a normal distribution. Even those students who hypothesize that income distribution is skewed may not realize the extent to which the median and mean incomes differ in the United States. This leads them to an inaccurate view of income distribution. This module prompts students to reconsider their perceptions of income inequality in the United States.

The Internal Revenue Service does not generally provide raw income tax return data. Thus, only aggregated data are available. The summary data include the number of tax returns in eighteen income categories or bins of unequal widths. In this module, students confront some of the problems in calculating summary statistics of data that have been grouped into bins of varying widths. Additional challenges include graphically representing data with a large range. Further complicating matters here is that the top income range is unbounded from above. In short, this data format makes it impossible to calculate the exact median and mean. In this module, students explore the challenges and limitations of calculating median and mean and representing the data visually for data reported in this way.

This module gives students an opportunity to think deeply about how to apply well-known ideas in a new setting. Students create and analyze histograms using their beliefs about income distribution and also data on income distribution. They explore ways to represent data that have been organized into income groups and frequencies. They approximate summary statistics of this grouped data, and they need to choose an upper bound for the top, unbounded income group. Furthermore, the data have a very large range and widely varying frequencies. Both of these make it difficult to represent the data with a histogram. This activity differs from the traditional setting where students work with a small set of data from which it is relatively straightforward to calculate the mean and median and produce a histogram from the data.

Income inequality has been and will continue to be an important issue. We created this lesson to provide students an opportunity to learn about income inequality in the United States and to experience the nuances of using aggregated data to tell the story of the data. One of the goals of this module is for students to gain an understanding of income distribution in the United States. After this lesson, we expect students to be able to

- describe the limitations of histograms for displaying data;
- describe the limitations of drawing conclusions from aggregated data;
- describe the level of income inequality in the United States;
- provide real-world justification for using the median instead of the mean to describe the center of a set of data.

1 Mathematical content

This module uses United States income data and is appropriate for any statistics course or a course that incorporates statistical reasoning.

Students should know how to construct a histogram and calculate the mean and median of a set of data. Students will work with real, aggregated data and experience the inherent challenges in describing grouped data in an exciting context. First, students cannot directly calculate the mean and median, so they must approximate these and explain how their assumptions about the distribution of the data within the bins affect their approximations. Second, the data are grouped in bins of unequal sizes, so students must confront their perceptions of the meaning of the heights of histogram bars. Third, the data range is so large that creating a single histogram that displays the data well is not possible. Fourth, the range for the uppermost income bracket is not bounded, so students must provide a reasonable bound that still allows one to represent the data graphically. Last, there is an optional opportunity for students to transform the data using a log transform (values provided) and explain how log transforming data affects the basic descriptive statistics of the data. In sum, students must decide how to apply familiar concepts in a novel context.

2 Context / Background

2.1 Income distribution. In 2011, the social justice movement known as *Occupy Wall Street* began protests that raised public awareness of what was perceived to be the unfair inequality in income distribution in the United States. One common measure of income inequality is the Gini coefficient (also called the Gini index). The Gini coefficient measures income inequality on a scale from 0 to 1, where a score of zero indicates that income is equally distributed across all individuals and a score of 1 means that one person acquires all of the income.[1] Depending on the data used to compute this coefficient, the Gini coefficient for the United States has ranged between 0.365 and 0.43 [**6**] or 0.469 and 0.555 [**18**] between 1979 and 2002. Differences in the range of these values can be attributed mainly to whether capital gains are included as income (inclusion increases the value). Regardless, the Gini coefficient was trending upwards during this time.

A special section in the May 23, 2014 issue of *Science* yielded topical commentary on income inequality throughout the world. Income distribution for the period 2008-2012 was more extreme in the United States than in Scandinavian countries but not as extreme as in Mexico or South Africa [**5**]. More generally, income inequality is much more pronounced in the emerging markets of developing nations than in high-income countries like the United States [**9**]. Income distribution is a complex statistic whose value depends on multiple factors (e.g., demography, education level, level of industrialization of the country, economic mobility). The Gini coefficient, for instance, does not incorporate economic mobility, and thus it should be noted that economic mobility between various income levels slightly reduces the effective magnitude of these numbers [**18**] since people do move between income classes. Regardless of the statistic that one might use to calculate income inequality, the data clearly show that income inequality is increasing [**16**].

2.2 Income data. The United States Internal Revenue Service (IRS) does not release all individuals' tax returns, so we must use available, aggregated data. We can obtain aggregated income data from the IRS, however, and provide these in Table 5.1 in Appendix A.

It is important to be aware of factors that may limit the strength of conclusions one may draw from the data. For example, using income tax data to derive time-series trends has been criticized for various reasons [**19**]; these shortcomings are addressed briefly in [**4**]. Furthermore, we expect a downward bias in reported income especially at the upper end of incomes. Top incomes may actually be larger than what is reported (due to evasion, compensation given in stock options or other non-wage formats, tax havens, etc.) [**4**]. Due to a minimum income of $0 and the possibility of some tax payers earning large incomes, household and personal income will be skewed right as is apparent by the large difference in median and mean incomes in Table 2.1.

2.3 Institutional background. We both work at Ithaca College, a primarily four-year undergraduate liberal arts and pre-professional college. We used this module in

[1] EDITOR'S NOTE: Chapter 9 ("Measures of Income Inequality" by Andrew J. Miller) in this volume explores the Gini index as well. Also see Chapter 15 ("Using Calculus to Model Income Inequality" by Bárbara González-Arévalo and Wilfredo Urbina-Romero) in the accompanying text *Mathematics for Social Justice: Resources for the College Classroom.*

Table 2.1. 2011 reference statistics on U.S. household income. The United States Census statistic is from `http://www.census.gov/prod/2013pubs/acsbr12-02.pdf` and the American Community Survey statistic is from `http://www.census.gov/acs/www/`.

Statistic	Value	Source
Median 2011 U.S. household income	$51324	United States Census
Mean 2011 U.S. household income	$70909	American Community Survey

the fourth week of class in an introductory statistics course titled *Statistics for Business, Economics, and Management*, for about 25 non-math majors. The typical student was a sophomore undergraduate majoring in business. Students sat at fixed tables, and the classroom included an instructor's workstation with a computer and an overhead projector. We projected overhead slides containing the sequence of prompts, discussion questions, and the histograms of the data. Students were encouraged to discuss many of their answers to questions with nearby students.

3 Instructor Preparation

3.1 Before the class meeting. Instructors should prepare one or several histograms of the data (for reference, we include some histograms in the appendix). The income groups and frequencies vary considerably, and the widths and relative heights of the bars are very disproportionate. Some software programs will not create a histogram from binned data, which may impact whether to ask the students to create a histogram using software.

Before class, instructors should also prepare photocopies of the data table. Some students may find it easier to approximate the mean and median of the income data with the data in hand.

Students should understand that histograms are different from bar charts. In bar charts, each bar has the same width. The relative heights of the individual bars reflect the relative data values. In histograms, the relative areas (not heights) of the bars must be proportional to the relative data values. Thus, histogram bars do not need to be of equal width. Since the income data are presented in groups of unequal sizes, the histogram bars will have unequal widths. Instructors should thoughtfully consider how to address this, especially if this will be the first time students will be seeing histograms with different bin widths.

3.2 During class. Instructors could consider beginning class with a discussion of student responses to the philosophical question

*"What income distribution is **fair**?"*

The discussion can be quite interesting and is likely to involve many of the students in the class.[2]

Students should receive the data only after they have completed sketching the first two histograms following the prompts which are provided in the next section (and in

[2] This question reminds us of a quote from the 1987 movie *Wall Street*, where a young stockbroker, Bud Fox, asks his boss Gordon Gecko, "How many yachts can you water-ski behind? How much is enough, huh?"

the sample handout found in Appendix B). One of the goals of this module is for students to gain an understanding of income distribution in the United States. This will be enhanced if their misbeliefs about income distribution are confronted at the beginning of this activity.

Once students have completed the first two histograms, some time should be allowed for a brief discussion of students' ideas. This is an opportunity to reinforce the differences in median and mean. It's also a great time to explain why data that has a hard lower (or upper) limit is commonly skewed (e.g., people's weights, the number of children in a family, test scores, etc.).

When students begin to create a histogram from the data (or just consider how to create such a histogram), they will need to choose (and justify) an upper bound for Adjusted Gross Income (AGI). One way to do this is to use a mathematical model from economics. If we assume that the upper tail of the income distribution follows a Pareto distribution[3], then the Pareto (or Pareto-Lorenz) coefficient for this model for 2011 is $\alpha = 1.60$ [**1**]. This implies that the mean income of individuals earning more than $10 million is $26.67 million[4]. If we assume that the mean income in each income bin is the midpoint of the bin, then the upper limit for the uppermost income bin would be $43.3 million.

4 The Module

The module can be used in one class or two consecutive classes. Students will do much of the work with partners or in small groups. Instructors may alternatively prefer to use a handout like the one in Appendix B to guide class work.

4.1 Discussion prompts.

The instructor can tailor the discussion questions below to suit the students' interests and the amount of time available.

"What income distribution is fair?" This open-ended question sets the tone for the module, letting students engage from the beginning of class. After students share with their neighbor, a brief discussion will bring some of these ideas forward and provide context for the next prompt.

"Sketch a histogram that represents what you believe to be the income distribution in the United States. Include your guesses for the mean and median incomes." Students will only need a couple of minutes for this. Instructors might also prompt students to describe the shape, center, and spread of the proposed income histogram.

"What do you think *should* be the income distribution in the United States? Include what the mean and median incomes should be." There are no wrong answers here. This question relates back to the question of fairness in how incomes are distributed. A follow-up prompt could be **"Provide justification for what you think the income distribution should look like."**

At this point, it's a good idea for students to share their histograms with a neighbor and explain what they drew and why. Pairs could be called upon to report what they

[3] See Section 5.3 Question 6.

[4] As explained on page 13 of [**4**], the mean income of individuals with income above $10 million will be $\beta(10,000,000)$ where $\beta = \frac{\alpha}{\alpha-1}$.

discovered (similarities and differences, assumptions, etc.) so that the big ideas can be acknowledged. Possible followup prompts could be: **"How many of you had a mean higher than the median? How many of you had a mean lower than the median? What is the general shape of your distribution if your mean was higher? lower?"**

4.2 Working with Data. At this point instructors can provide the students with the income data table (see Appendix A) and ask them what they find interesting. This may be the first time that students have heard the word "bin" in the context of data or a histogram, so this term might need to be defined. Perhaps it is relevant to have a discussion about why the top income group is unbounded. Then, students should be prompted to create a histogram from the income data from the column "Number of total returns" in Table 5.1. An alternative to having students attempt to create the histogram is to provide slides of the histograms provided in Appendix A (see Figures 6.1-6.3). Since the bin corresponding to the greatest incomes does not have an upper limit, students should choose their upper limit thoughtfully, and justify their choices. A follow-up question is **"How do we deal with the uppermost income bin being unbounded?"**

After discussing how one might choose an upper limit, students then should estimate the mean and median incomes. Getting different ideas from the class provides an opportunity to clear up any misconceptions. Then, the instructor can ask students to compare the means and medians of the two histograms they sketched and compare these values with their estimates from the table. A good followup question to ask is **"Which of these statistics (mean or median) is a better representation of the "average" household income?"** Again, students can compare with their neighbor and report back to the class.

Assumptions abound in statistics, so it is appropriate to ask the students **"What assumptions did you need to make to estimate the mean and median? Justify why these are reasonable assumptions."** A follow-up question is **"How does using the lower, midpoint, or upper points of each bin interval affect the mean and median incomes?"**

Once students understand the income data histogram, it is time to have them compare this histogram with the other two histograms they sketched. Suitable prompts could be: **"How are they the same? How do they differ? Which of these histograms is most skewed? How did you arrive at your choice?"**

One way to wrap up the activity is to ask students **"What do these histograms tell us about income inequality in the United States? What surprises you about income inequality in the United States?"**

Optional Activity: As an optional activity, instructors can have the students create a histogram using the log transformed data from Table 5.2. Students can then compare the mean and median of the log transformed data. Possible questions are **"Why might someone want to log transform a dataset? What are the benefits/drawbacks?"**

4.3 Additional discussion questions. Below we list some discussion questions that would encourage students to delve deeper into the issue at hand:

(1) *How does the presentation of data in bins affect our calculating the mean and median of the data? Does it make histograms easier/harder to create? What, if anything, do*

we gain (and what do we lose) from the presentation of data in this form versus the raw data? Who gets to choose the bin sizes?

(2) *The median U.S. household income is $51324* [**15**], *and the mean U.S. household income is $70909* [**2**]. *Why might these be higher than the Adjusted Gross Income (AGI) data that we have? What might we hypothesize about the income distribution of households? (more skewed, less skewed, and why)*

(3) *Income data can come from many sources, e.g., tax authorities or surveys. What are the pros and cons of the conclusions one can draw from these various sources?*

(4) *Equality of opportunity in the United States does vary based on geography* [**14**]. *Additionally, the Gini coefficient varies from state to state. Which state do you think would have the greatest income inequality? Least? Why do you think so (think about the factors that might contribute to income inequality)? Look over the list in Table* 5.3. *Are there any surprises? What hypotheses might you consider testing after seeing this data?*

(5) *Use the additional data from Table* 5.1 *to compare income distributions (e.g., married filing jointly vs. single).*

As an example of the many directions these discussions may take, we offer below a few possible answers to Question (3):

- Tax data are the most reliable, but U.S. personal tax returns are private and confidential so raw tax data is unavailable from the Internal Revenue Service. Tax laws change over time and can cause AGI calculations to change over time so that time series may be inappropriate unless adjustments are made.
- Survey data are more likely to under-represent income (although Darrell Huff's classic book *How to Lie with Statistics* [**8**] has an interesting story about Yale alumni over-reporting income).
- Non-response and small sample sizes may skew survey income results downward.

5 Additional Thoughts

5.1 Application notes. When we used this module, we started by asking our students "**What income distribution is fair?**" Students shared their ideas with a partner. Then, we asked them to identify the important ideas that came up. Students sketched the first two histograms. They were prompted to discuss with the person next to them the assumptions that went into the sketches. Then, as a class, students identified the common big ideas. Students saw a slide of the data table and were prompted, "**What's interesting?**" Then, they were asked to compare their histograms with the data table. This took about twenty-five minutes which was all the time remaining in that initial class period.

The second day continued with a similar pedagogy: prompts, thinking, sharing, and reporting. We gave students a handout with the data table, but we did not provide any other handouts. The prompts and questions from Part Two (see module handout in Appendix B) were presented again on an overhead screen. Next time, we would consider putting the questions listed in Part Two on a worksheet for students to consider in small groups. This would be followed by a whole-class reporting and discussion. Students pointed out that the majority of people's income was lower than they expected. They learned that more than half of the population is earning less than $40,000 (see the tabular data to find that the median occurs in the $30,000-$39,999 income bracket). They remarked that there are some extreme outliers that create a higher mean than median and that income inequality is far more skewed to the right than expected.

Students noticed that approximating the median was much easier than approximating the mean. Those students who had taken calculus brought up the topic of Riemann sums during the discussion on calculating the mean income. Some students remarked that the level of income inequality was not good while another student mentioned that it was a lot different from what he thought it was. He thought there would be more people in higher income brackets. Students enjoyed comparing the mean and median of U.S. household income with their adjusted gross income data. They were surprised at the effect of deductions on income.

Towards the conclusion of the class period, students were asked, "**What more would you like to explore about this, or a related, topic?**" Their answers indicated an awareness of the importance of asking questions of the data collection method, methods for working with data reported in ranges, and real-world quantitative inquiries about income and wealth.

Representative student responses included

- *I would like to learn about any data that may have been left out.*
- *I would like to know more about ways to find the average when given data in ranges.*
- *I would like to know more about income distribution by regions.*
- *I would like to learn more about finding more exact numbers to get more accurate estimates.*
- *I would like to know more about the amount of wealth instead of income, and see how the curve changes.*
- *Why do the majority of people have such low incomes?*
- *Is there a way to properly distribute the wealth so that you deserve the income you receive?*
- *How does this relate to the 2016 election?*
- *How do inflation rates affect this data set?*
- *At what rates are people in these income brackets taxed?*
- *What would the data look like in smaller brackets and before deductions?*

Overall, we used one and a half fifty-minute class periods for this module. We had time to consider the first two questions from Section 4.3, but did not have time to pursue any of the questions from Section 5.3 as we had originally planned.

5.2 Possible variations in implementation. Students do not need to create the histogram from the data. They may encounter large difficulties in working with a dataset in this format. The instructor may choose to allow students to attempt to create the histogram which could lead to a discussion about the challenges in creating such a histogram. This may lead to an idea that perhaps a tabular representation of the data is best. Or, the instructor may post pictures of histograms that attempt to display the data. A discussion about these histograms could follow.

In Appendix B, there is a student handout containing the instructor prompts. Rather than present the module as a series of prompts and responses, the instructor may choose to photocopy the handout and provide this to students as an in-class worksheet. Alternatively, parts of the worksheet could be completed for homework.

5.3 Extensions. Below we list a few additional lines of inquiry which could supplement the module.

(1) Additional income data on married couples filing jointly and singles is found in Table 5.1. The instructor can use these for homework reinforcement of the concepts in this module.

(2) State-by-state Gini coefficient data, provided in Table 5.3, provides opportunities to explore a multitude of questions. For example: (1) **Is the Gini coefficient higher or lower in states that have traditionally had a Republican or Democratic governor?**; (2) **Does the Gini coefficient show trends among, for example, more populated states or regions of the United States?**; (3) **Does the Gini coefficient depend on the heterogeneity in education level? in the per capita welfare rate?**

(3) The Statistics of Income (SOI) division of the Internal Revenue Service issues reports on income distribution in the United States; see [**12**]. Conference papers on individual tax statistics are listed at [**13**]. These might be good resources for further work.

(4) After the initial discussion, students may be interested in watching the video *Wealth Inequality in America* [**17**] (or reading [**3**]) which highlights the distribution of wealth in the United States. Also of possible interest is [**7**], an article about this video. Instructors planning to use this video should be prepared to explain how one could create a graph similar to the one in the video from the data in this module.

(5) If students watch the video *Wealth Inequality in America* [**17**] about wealth inequality in the United States, it can lead to a discussion, assignment, or project about the differences between wealth and income. Some followup questions that may trigger a good discussion are: **Is it true that those with much wealth are also the top earners? How could one measure the amount of economic mobility (movement between income classes)?** Instructors might find [**18**] relevant here.

(6) Instructors teaching a calculus-based statistics course might consider having students use the Pareto distribution to derive the fact that the ratio between the average income of individuals with income above y^* to the cumulative income that is above this value y^* is constant. This allows us to estimate the average income of any income bin with an unbounded upper limit.

The cumulative density function of the Pareto distribution is given by

$$F(y) = 1 - \left(\frac{k}{y}\right)^{\alpha}.$$

From this, students can derive the probability density function

$$F'(y) = f(y) = \frac{\alpha}{y}\left(\frac{k}{y}\right)^{\alpha}.$$

The total income for the proportion of individuals with income above y^* is given by $\int_{y^*}^{\infty} zf(z)\,dz$, and the total proportion of people whose incomes are above y^* is given by $\int_{y^*}^{\infty} f(z)\,dz$; therefore, the mean income of individuals with income above y^* is given by the ratio of these. The derivation, available in [4], is straightforward, and is included below:

$$\frac{\int_{y^*}^{\infty} zf(z)\,dz}{\int_{y^*}^{\infty} f(z)\,dz} = \frac{\int_{y^*}^{\infty} z\frac{\alpha}{z}\left(\frac{k}{z}\right)^{\alpha} dz}{\int_{y^*}^{\infty} \frac{\alpha}{z}\left(\frac{k}{z}\right)^{\alpha} dz} = \frac{\alpha k^{\alpha}\int_{y^*}^{\infty} z^{-\alpha}\,dz}{\alpha k^{\alpha}\int_{y^*}^{\infty} z^{-\alpha-1}\,dz}$$

$$= \frac{z^{-\alpha+1}|_{y^*}^{\infty}}{z^{-\alpha}|_{y^*}^{\infty}}\cdot\frac{-\alpha}{-\alpha+1} = \frac{0-(y^*)^{-\alpha+1}}{0-(y^*)^{-\alpha}}\cdot\frac{\alpha}{\alpha-1}$$

$$= y^*\left(\frac{\alpha}{\alpha-1}\right)$$

We see that the ratio of the mean income above y^* to the income level y^* is constant and does not depend on y^*. Thus, if $\alpha = 1.6$, then $\frac{\alpha}{\alpha-1} = 2.667$, and according to this model, the average income of individuals with income over 10 million would be \$26.667 million. Similarly, the average income of individuals with income over 1 million would be \$2.6667 million.

Bibliography

[1] Facundo Alvaredo, Anthony B. Atkinson, Thomas Piketty, Emmanuel Saez, and Gabriel Zucman. *WID-The World Wealth and Income Database*, http://www.wid.world/, accessed on September 18, 2016.

[2] American Community Survey, from `http://www.census.gov/programs-surveys/acs/`, accessed on August 21, 2016.

[3] Dan Ariely, "Americans Want to Live in a Much More Equal Country (They Just Don't Realize It)." *The Atlantic*, August 2, 2012.

[4] A. B. Atkinson, T. Piketty, and E. Saez. "Top incomes in the long run of history." *Journal of Economic Literature*, **49** (2011), 3–71.

[5] G. Chin and E. Culotta. "The science of inequality." *Science*, **344** (2014), 819–821.

[6] E. Gudrais. "Unequal America." *Harvard Magazine*, July-August 2008, from `http://harvardmagazine.com/2008/07/unequal-america-html`, accessed on August 21, 2016.

[7] J. Harkinson, "The Great American Inequality Video." *Mother Jones*, March 4, 2013, from `http://www.motherjones.com/mojo/2013/03/great-american-inequality-video`, accessed on August 21, 2016.

[8] D. Huff. *How to Lie with Statistics*. New York: Norton, 1993.

[9] M. Hvistendahl. "While emerging economies boom, equality goes bust." *Science*, **344** (2014), 832–835.

[10] Internal Revenue Service, "Definition of adjusted gross income," from `http://www.irs.gov/uac/Definition-of-Adjusted-Gross-Income`, accessed on August 21, 2016.

[11] Theodore S. Sims, *Income taxation, wealth effects, and uncertainty: portfolio adjustments with isoelastic utility and discrete probability*, Econom. Lett. **135** (2015), 52–54, DOI 10.1016/j.econlet.2015.07.006. MR3398812

[12] Internal Reve
Stats-Archi

[13] Internal Reve
https://www
statistics,

[14] J. Mervis. "Tr

[15] A. Noss. "Hou
gov/library

[16] T. Piketty and

[17] politizane, *W*
cessed on Au

[18] M. Strudler, T
2002." Intern
accessed on A

[19] R. M. Titmus

[20] United States
from https:/
July 22, 2014.

[21] D. H. Weinb
States Censu
21, 2016.

[12] Internal Revenue Service, "SOI Tax Stats – Archive," from https://www.irs.gov/uac/SOI-Tax-Stats-Archive, accessed on February 24, 2016,

[13] Internal Revenue Service, "SOI Tax Stats – Conference Papers on Individual Tax Statistics," from https://www.irs.gov/uac/soi-tax-stats-conference-papers-on-individual-tax-statistics, accessed on February 24, 2016.

[14] J. Mervis. "Tracking who climbs up – and who falls down – the ladder." *Science*, **344** (2014), 836–837.

[15] A. Noss. "Household income: 2012", United States Census Bureau, available at https://www.census.gov/library/publications/2013/acs/acsbr12-02.html, accessed on August 21, 2016.

[16] T. Piketty and E. Saez. "Inequality in the long run." *Science*, **344** (2014), 838–843.

[17] politizane, *Wealth Inequality in America*, video available at https://youtu.be/QPKKQnijnsM, accessed on August 21, 2016.

[18] M. Strudler, T. Petska, and R. Petska. "Further analysis of the distribution of income and taxes, 1979-2002." Internal Revenue Service, available at https://www.irs.gov/pub/irs-soi/04asastr.pdf, accessed on August 21, 2016.

[19] R. M. Titmuss. *Income distribution and social change.* London: Allen & Unwin, 1962.

[20] United States Census Bureau, "Gini ratios for households, by race and Hispanic origin of householder", from https://www.census.gov/hhes/www/income/data/historical/inequality/, accessed on July 22, 2014.

[21] D. H. Weinberg. "United States neighborhood income inequality in the 2005-2009 period," United States Census Bureau, http://www.census.gov/prod/2011pubs/acs-16.pdf, accessed on August 21, 2016.

Appendix A: Tables and Datasets

Table 5.1. 2011 Adjusted Gross Income (AGI) for all types of returns (married filing jointly, married filing separately, head of household, and surviving spouse) and two categories broken out: married filing jointly and singles [**11**]. Adjusted Gross Income is defined as gross income minus adjustments to income [**10**].

Adjusted Gross Income	Number of total returns	Married filing jointly	Single persons
$1-$4,999	10,692,838	762,132	9,173,578
$5,000 - $9,999	12,386,716	1,200,915	9,104,803
$10,000 - $14,999	12,925,831	1,763,357	7,770,614
$15,000 - $19,999	11,880,059	2,037,870	6,451,097
$20,000 - $24,999	10,210,706	2,171,240	5,278,728
$25,000 - $29,999	8,987,613	2,122,016	4,473,611
$30,000 - $39,999	14,520,079	4,070,283	7,032,475
$40,000 - $49,999	10,983,973	3,974,111	4,973,871
$50,000 - $74,999	18,949,278	9,774,323	6,592,040
$75,000 - $99,999	11,926,401	8,512,981	2,517,798
$100,000 - $199,999	14,755,766	12,229,408	1,865,403
$200,000 - $499,999	3,801,641	3,278,621	397,964
$500,000 - $999,999	597,525	512,619	63,675
$1,000,000 - $1,499,999	134,907	115,095	14,929
$1,500,000 - $1,999,999	55,986	46,567	6,854
$2,000,000 - $4,999,999	79,363	65,272	9,996
$5,000,000 - $10,000,000	19,189	15,557	2,501
$10,000,000+	11,445	9,127	1,479

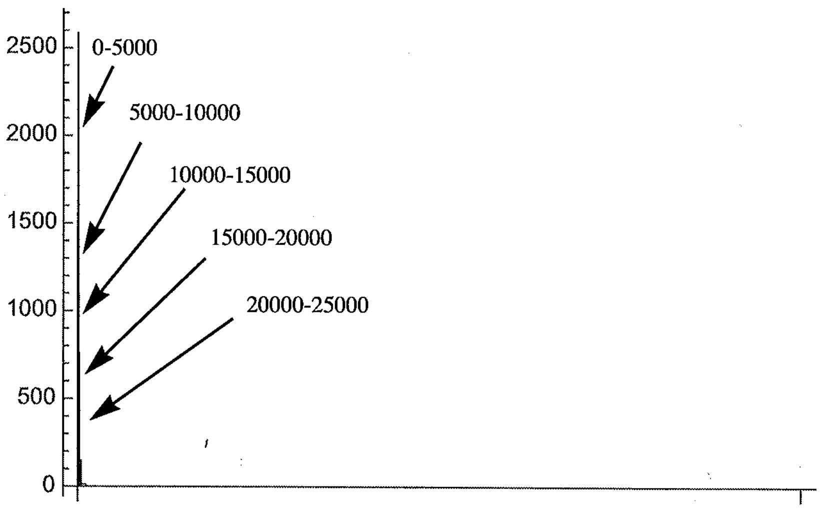

Figure 6.1. Histogram of the number of total returns from Table 5.1.

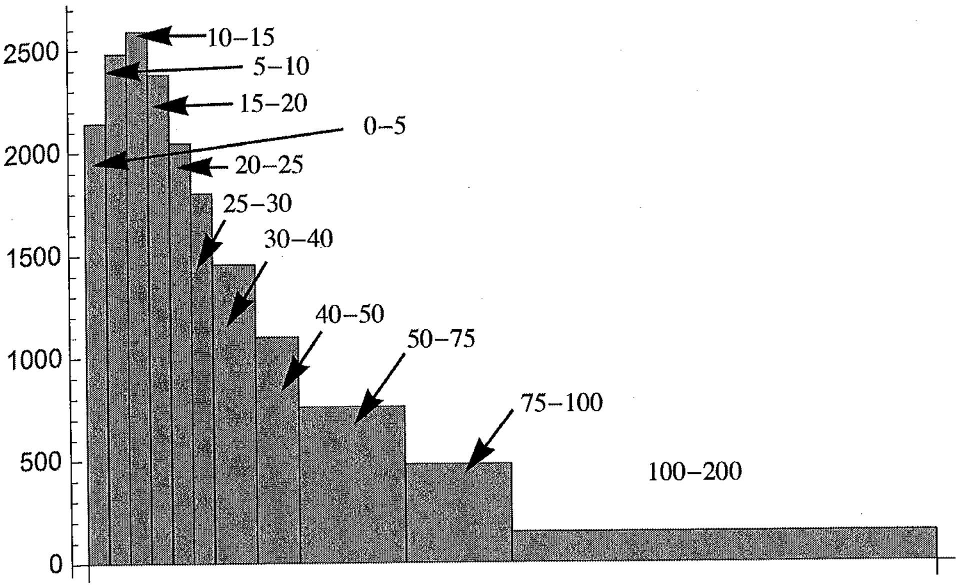

Figure 6.2. Histogram of the number of total returns for the lowest eleven income brackets from Table 5.1. The height of each bar is the density (number of returns in the income bracket / $ width of the income bracket).

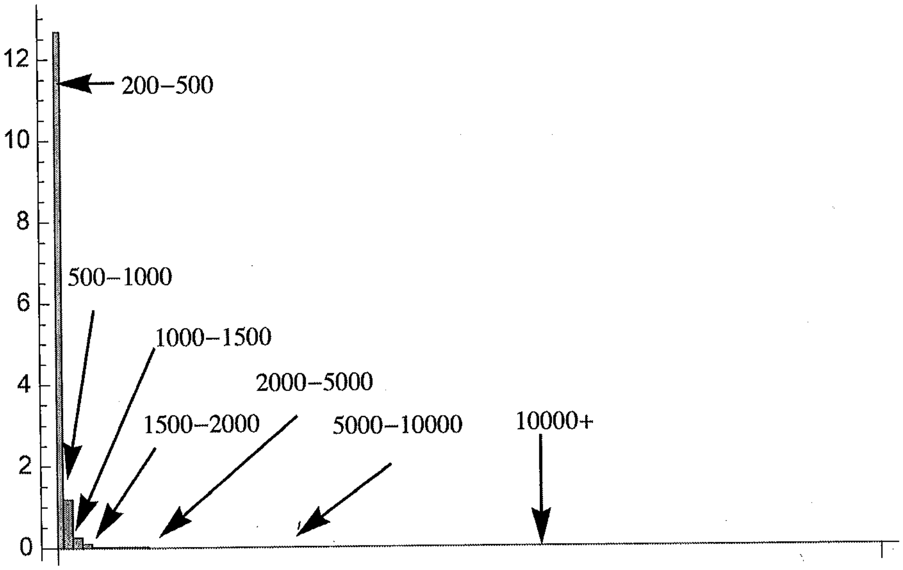

Figure 6.3. Histogram of the number of total returns for the highest seven income brackets from Table 5.1. The upper bracket has been closed at $43.3 million. The height of each bar is the density (number of returns in the income bracket/$ width of the income bracket).

Table 5.2. 2011 Tax year, log transformed data

Log (base 10) Adjusted Gross Income	Number of total returns
$0.0000 - $3.6989	10,692,838
$3.6990 - $4.0000	12,386,716
$4.0000 - $4.1761	12,925,831
$4.1761 - $4.3010	11,880,059
$4.3010 - $4.3979	10,210,706
$4.3979 - $4.4771	8,987,613
$4.4771 - $4.6020	14,520,079
$4.6020 - $4.6990	10,983,973
$4.6990 - $4.8751	18,949,278
$4.8751 - $5.0000	11,926,401
$5.0000 - $5.3010	14,755,766
$5.3010 - $5.6990	3,801,641
$5.6990 - $6.0000	597,525
$6.0000 - $6.1761	134,907
$6.1761 - $6.3010	55,986
$6.3010 - $6.6990	79,363
$6.6990 - $7.0000	19,189
$7.0000+	11,445

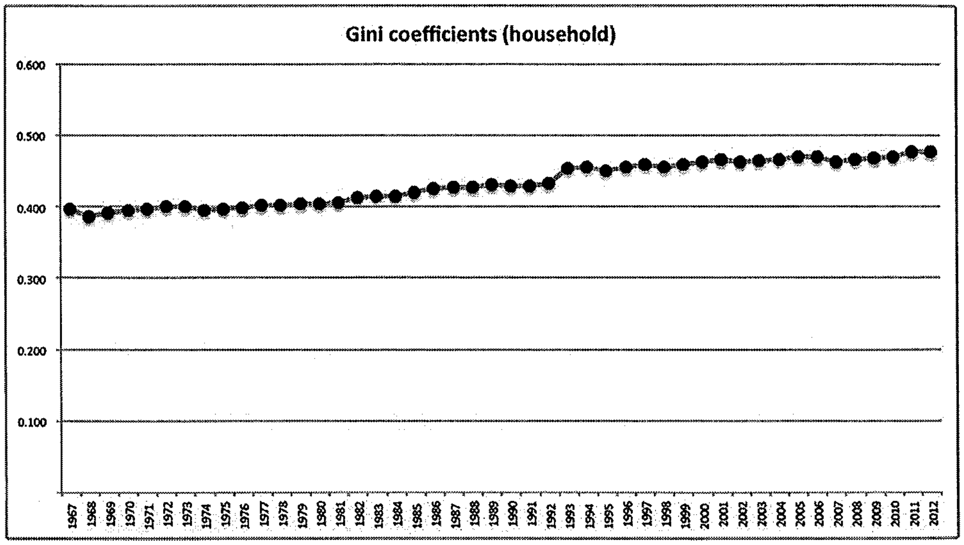

Figure 6.4. Gini coefficients for households [20].

Appendix B: Assignments and Handouts

Student Worksheet: Income Distribution in the United States. Income inequality exists in any economic system where income varies from person-to-person. In other words, it is pervasive and unavoidable. The Gini coefficient measures income inequality on a scale from 0 (no inequality) to 1 (extreme inequality). What can **you** say about income inequality from Figure 6.4?

Table 5
(includ
2005-20

St
District of
New
Conn
Loui
Miss
Te
Ala
Cali
Fl
Ten
Unit
Ge
Il
Mass
North
Ke
Ne
South
Ar
Ok
Ne
Pen
V
We
Rh
G

What do
question pro
data to answ
before you k
you collect t

In Part O
the United S
tribution his
histograms.
tremely pers
known only

Part One:

(1) *Sketch a*
the Uni
the sha

(2) *Repeat*
should

Table 5.3. Gini coefficients for household income inequality by state (including Washington, D. C. and the United States, for reference): 2005-2009 [**21**].

State	Gini coefficient	State	Gini coefficient
District of Columbia	0.540	Arizona	0.451
New York	0.499	Missouri	0.449
Connecticut	0.481	Michigan	0.448
Louisiana	0.477	Ohio	0.447
Mississippi	0.474	Oregon	0.447
Texas	0.474	North Dakota	0.443
Alabama	0.471	Kansas	0.442
California	0.469	Washington	0.442
Florida	0.469	Maryland	0.439
Tennessee	0.468	South Dakota	0.439
United States	0.467	Montana	0.438
Georgia	0.465	Delaware	0.435
Illinois	0.465	Minnesota	0.435
Massachusetts	0.465	Maine	0.434
North Carolina	0.463	Indiana	0.433
Kentucky	0.462	Nevada	0.433
New Jersey	0.462	Nebraska	0.430
South Carolina	0.460	Vermont	0.430
Arkansas	0.459	Hawaii	0.427
Oklahoma	0.459	Idaho	0.426
New Mexico	0.458	Iowa	0.426
Pennsylvania	0.457	Wisconsin	0.426
Virginia	0.455	Wyoming	0.424
West Virginia	0.453	New Hampshire	0.418
Rhode Island	0.452	Alaska	0.411
Colorado	0.452	Utah	0.411

What does income inequality look like in the United States today? To answer this question properly, we need data. Nonetheless, part of the fun and the science of using data to answer a question is formulating your best guess as to what the answer will be before you know anything about the data (and in most cases, this occurs long before you collect the data).

In Part One, you will sketch two hypothetical histograms of income distribution in the United States. Afterwards, you will use the data in Part Two to create an income distribution histogram. As you complete your work, take some time to compare the three histograms. Data from a reliable source which are well-represented visually can be extremely persuasive and help to tell a story. The complete story, however, may become known only after other questions have been asked and answered.

Part One: Make a hypothesis.

(1) *Sketch a histogram that represents what you believe to be the income distribution in the United States. Include your guesses for the mean and median incomes. Describe the shape, center, and spread of your proposed income histogram.*

(2) *Repeat this exercise by sketching a second histogram representing what you think should be the income distribution in the United States. Include approximations of*

the mean and median incomes, and describe the shape, center, and spread of this histogram. Give justification for what you think the income distribution should look like.

Part Two: Use the data.

(1) *Create a histogram from the income data from the column "Number of total returns" in Table* 5.1. *Notice that the bin corresponding to the greatest incomes does not have an upper limit. Thoughtfully choose an upper limit, and justify your choice.*

(2) *Estimate the mean and median incomes. How do they compare, and why?*

(3) *Which of these statistics is a better representation of the "average" household income? What assumptions did you need to make to estimate the mean and median? Justify why these are reasonable assumptions.*

(4) *Compare this histogram with the other two. How are they the same? How do they differ? Which of these histograms is most skewed? How did you arrive at your choice?*

(5) *What do these histograms tell us about income inequality in the United States? What surprises you about income inequality in the United States?*

*6. *Repeat the previous steps using the log transformed data from Table* 5.2. *Compare the mean and median of the log transformed data.*

*7. *Why might someone want to log transform a dataset? What are the benefits/ drawbacks?*

* – optional