STAT 113
Describing Quantitative Data: Shape and Center

Colin Dawson

September 12, 2017

Histogram: Smoking During Pregnancy

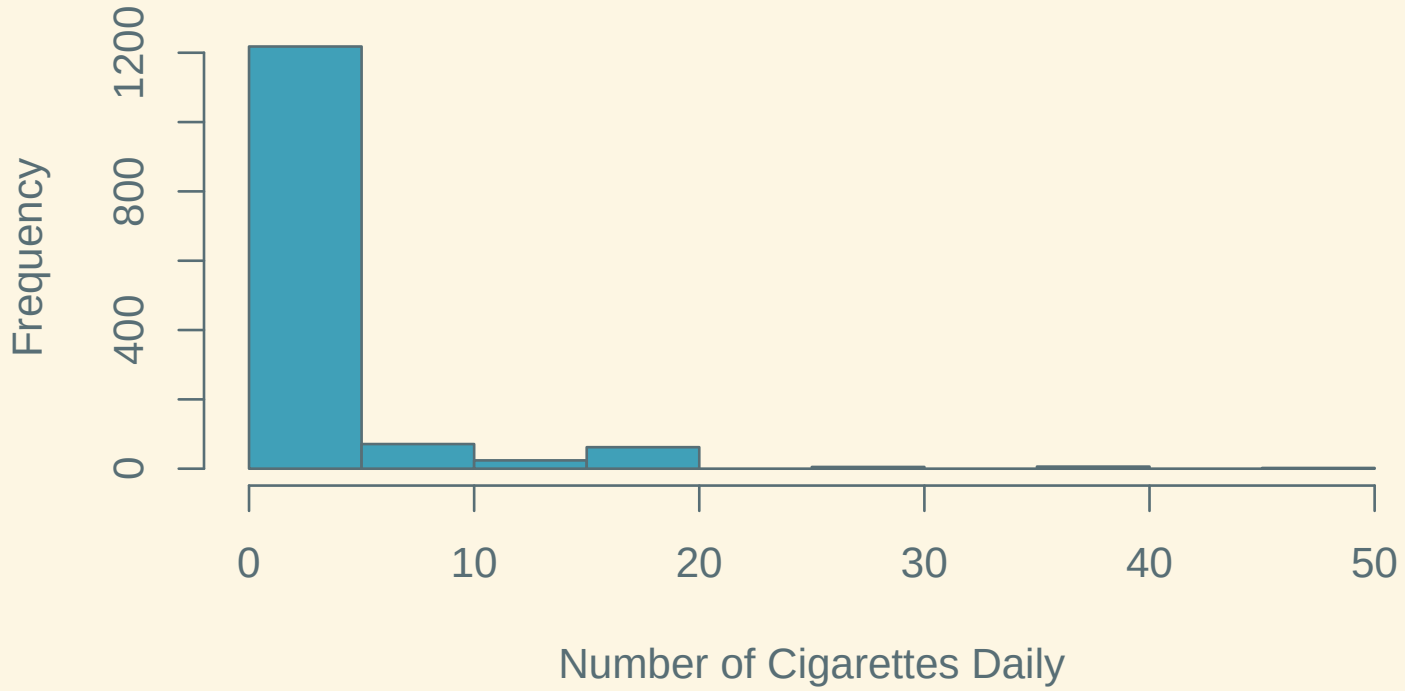


Figure: Daily cigarettes smoked by pregnant women (Source: Wooldridge, 2000)

What do you think of this histogram?

Example: Smoking During Pregnancy

- Here's a frequency table...

Cigarettes	Frequency
0	1176
1	3
2	4
3	7
4	9
5	19
6	6
7	4
8	5

Cigarettes	Frequency
9	1
10	55
12	5
15	19
20	62
30	5
40	6
46	1
50	1

Example: Smoking During Pregnancy

What can we do to improve the presentation of the data?

- Remove nonsmokers; present smokers vs. non- in a separate chart
- We still have a problem with uneven precision.
- Could be better to use a categorical variable (unequal ranges) instead

A Histogram With Variable Width Bins

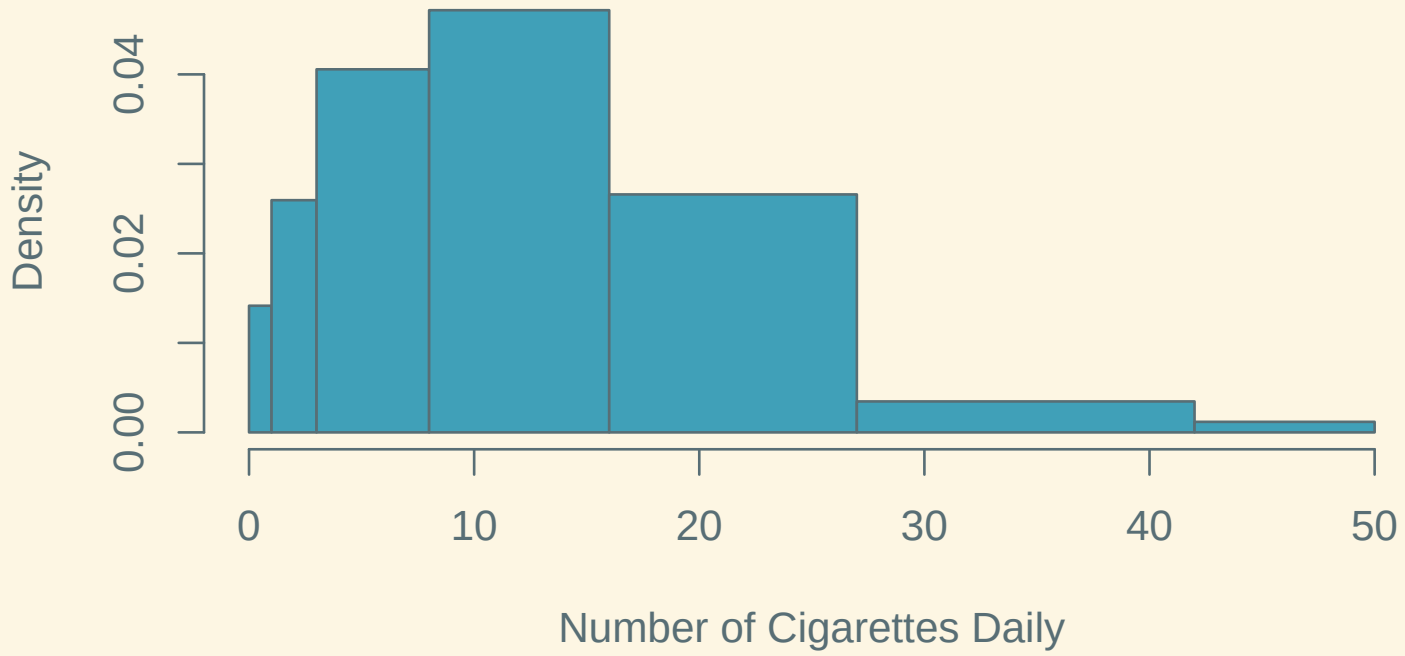


Figure: Daily cigarettes smoked by pregnant women (Source: Wooldridge, 2000). Shown: women who smoke at least 1 cigarette daily (212 out of 1388 surveyed)

Narrowing Bins: Frequency

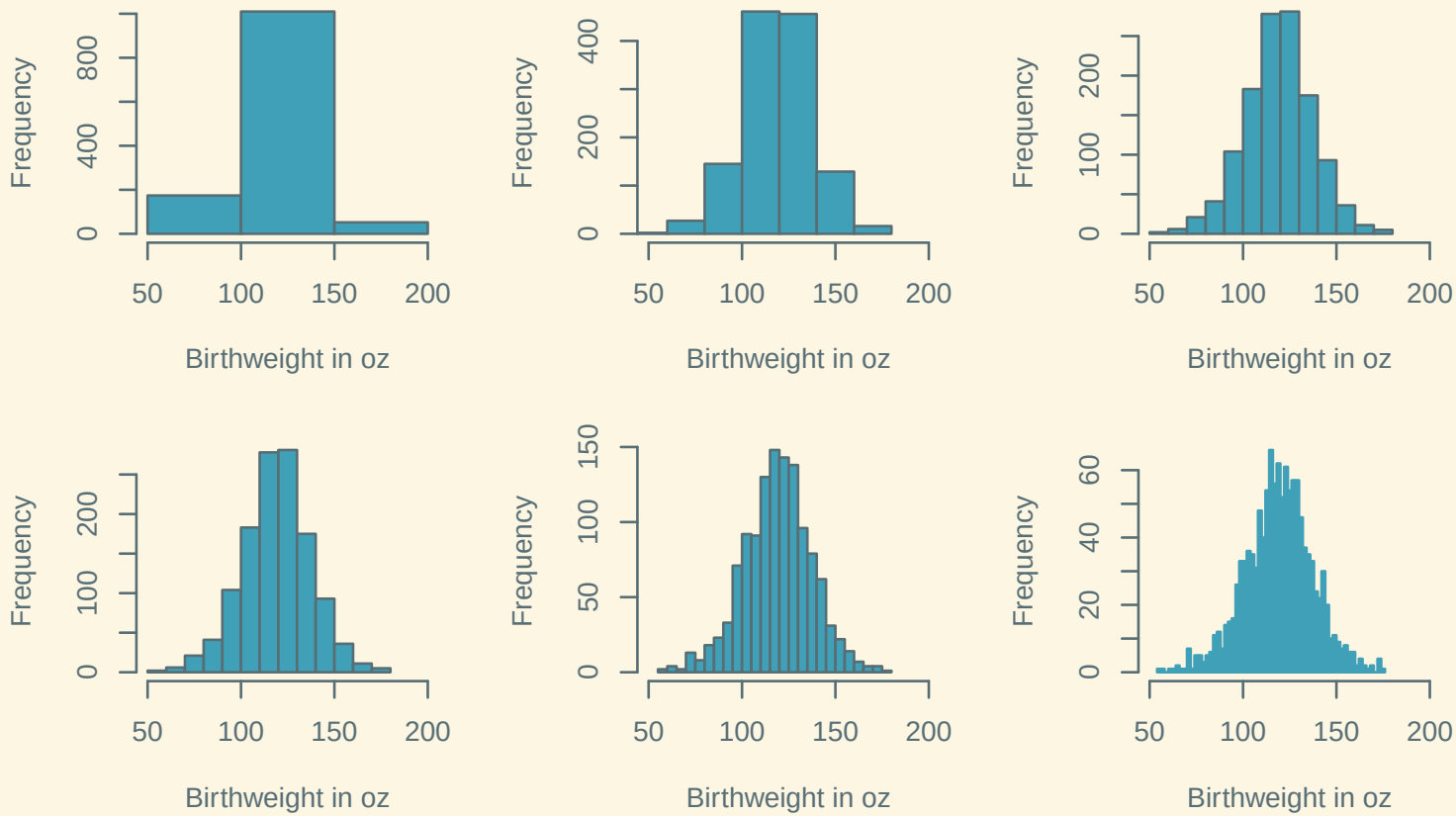


Figure: Histograms of Babies' Birth Weights (Nolan and Speed, 2000)

Density Functions

- With a continuous variable, the "edges" between bins are artificial.
- If we keep collecting data, expect the histogram to smooth out.
- Can capture what might happen with more data using an (estimated) **density function**: smooth curve showing the shape of the data distribution. (Extrapolation to "infinite" data.)

Density Functions

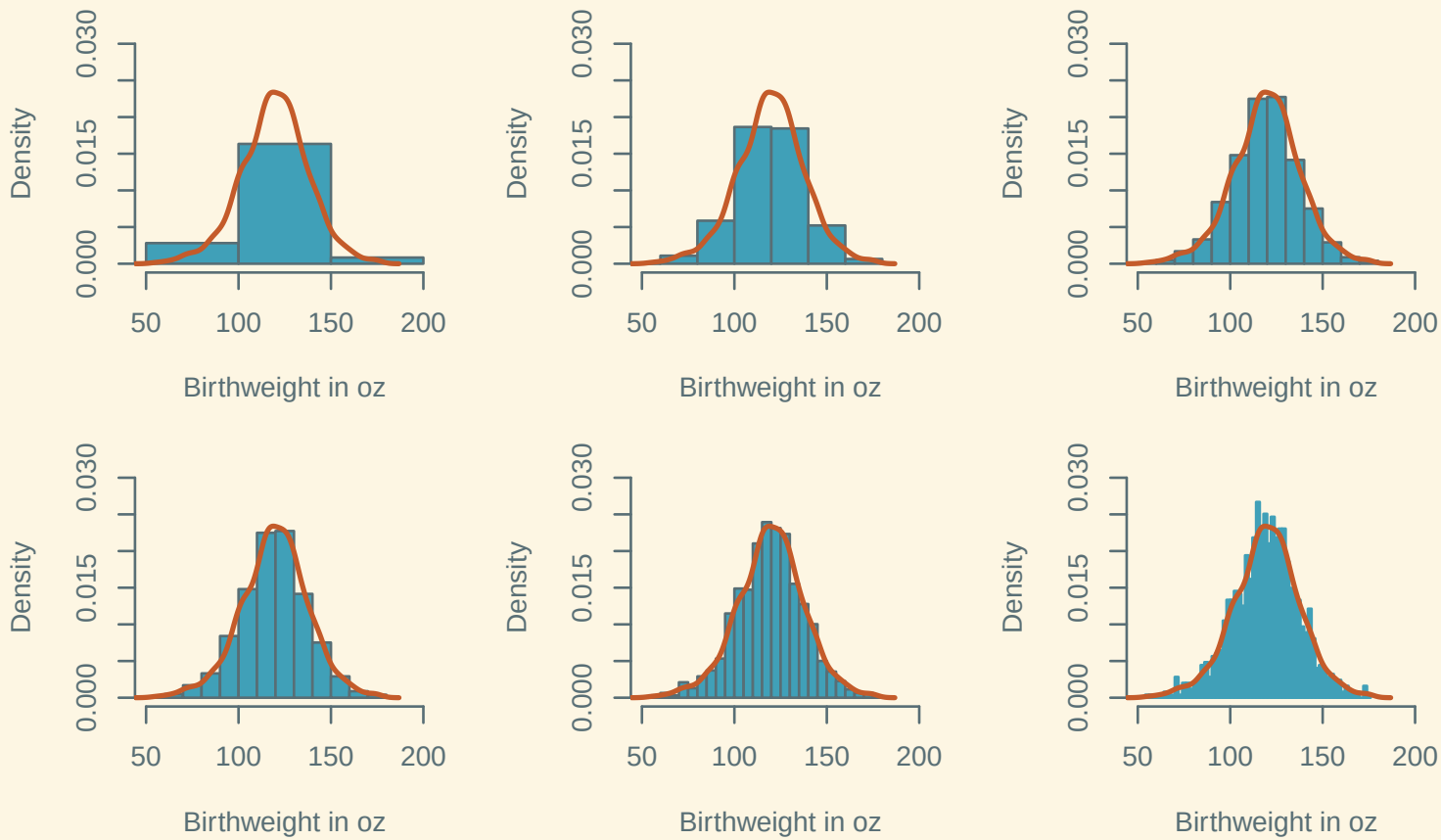


Figure: Densities of Babies' Birth Weights (Nolan and Speed, 2000)

Distribution of a Quantitative Variable

The distribution of a quantitative variable is characterized by:

A. Shape (symmetric, skewed, bimodal, etc.)
B. Center (mean, median)
C. Spread (Interquartile Range, Standard Deviation)
D. Outliers (if any)

Distribution of a Quantitative Variable

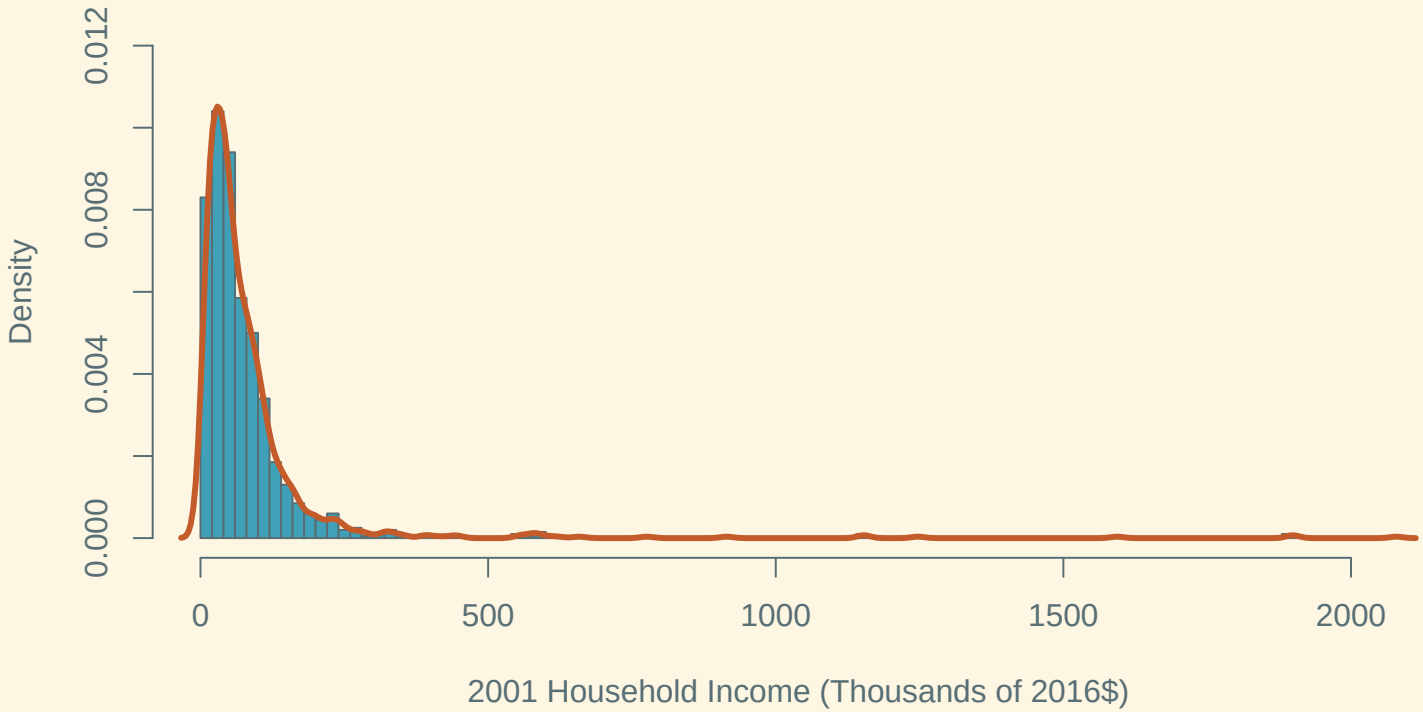
The distribution of a quantitative variable is characterized by:

A. **Shape (symmetric, skewed, bimodal, etc.)**
B. Center (mean, median)
C. Spread (Interquartile Range, Standard Deviation)
D. Outliers (if any)

Skewness

- A distribution is **skewed** when the extreme values on one side are more extreme than those on the other.
- We call a distribution **right-skewed** when the longer "tail" is on the right, and **left-skewed** when the longer tail is on the left.

Right Skew


Density
0.000
0.004
0.008
0.012
0
500
1000
1500
2000
2001 Household Income (Thousands of 2016$)

Distribution of a Quantitative Variable

The distribution of a quantitative variable is characterized by:

A. Shape (symmetric, skewed, bimodal, etc.)
B. **Center (mean, median)**
C. Spread (Interquartile Range, Standard Deviation)
D. Outliers (if any)

Central Tendency

- Often we want to summarize data with a single number
- Usually representing a "typical", or "middle" value.
- But how do we define "typical"?
- Depends on the data and the question.

Baby Weights

Where's a typical value?

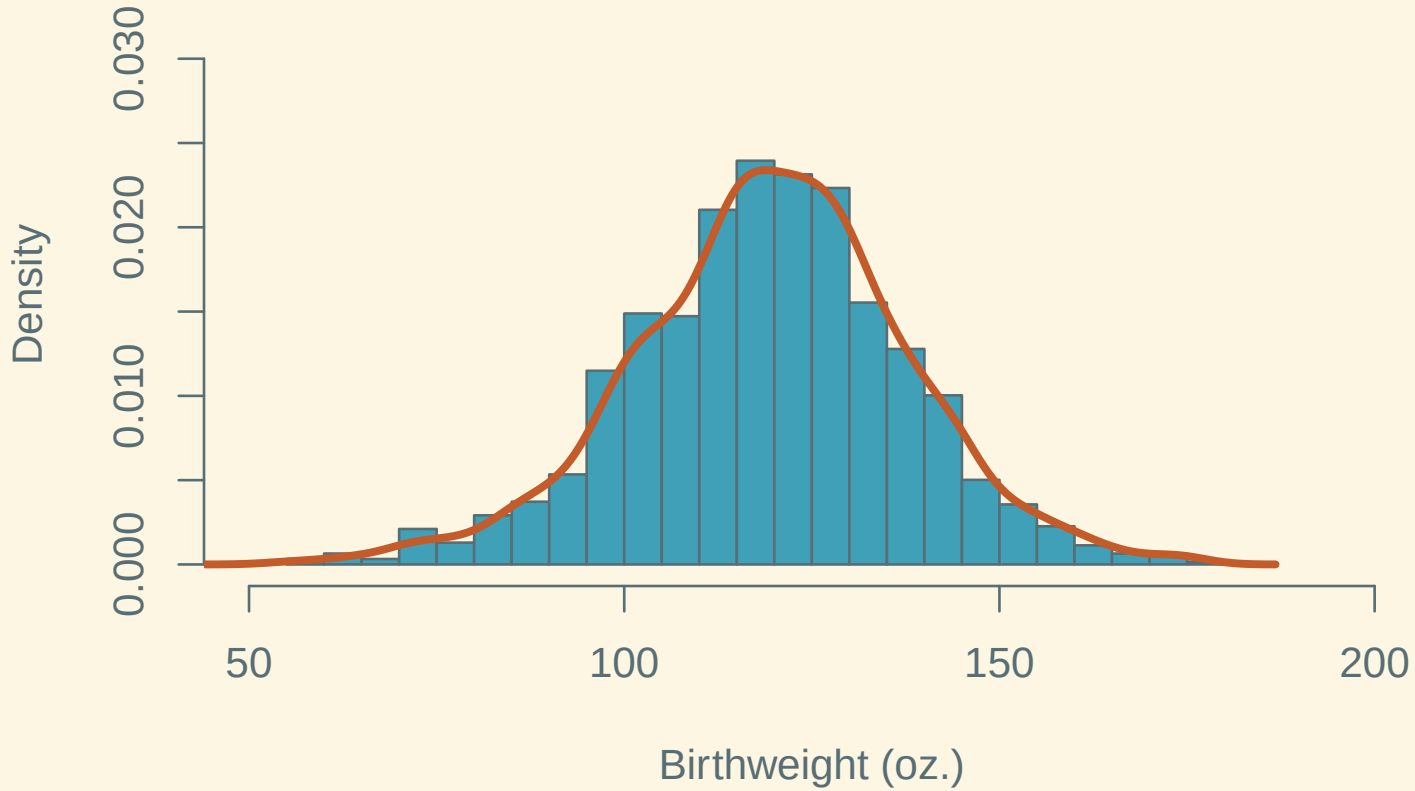


Figure: Distribution of Babies' Birth Weights (Nolan and Speed, 2000)

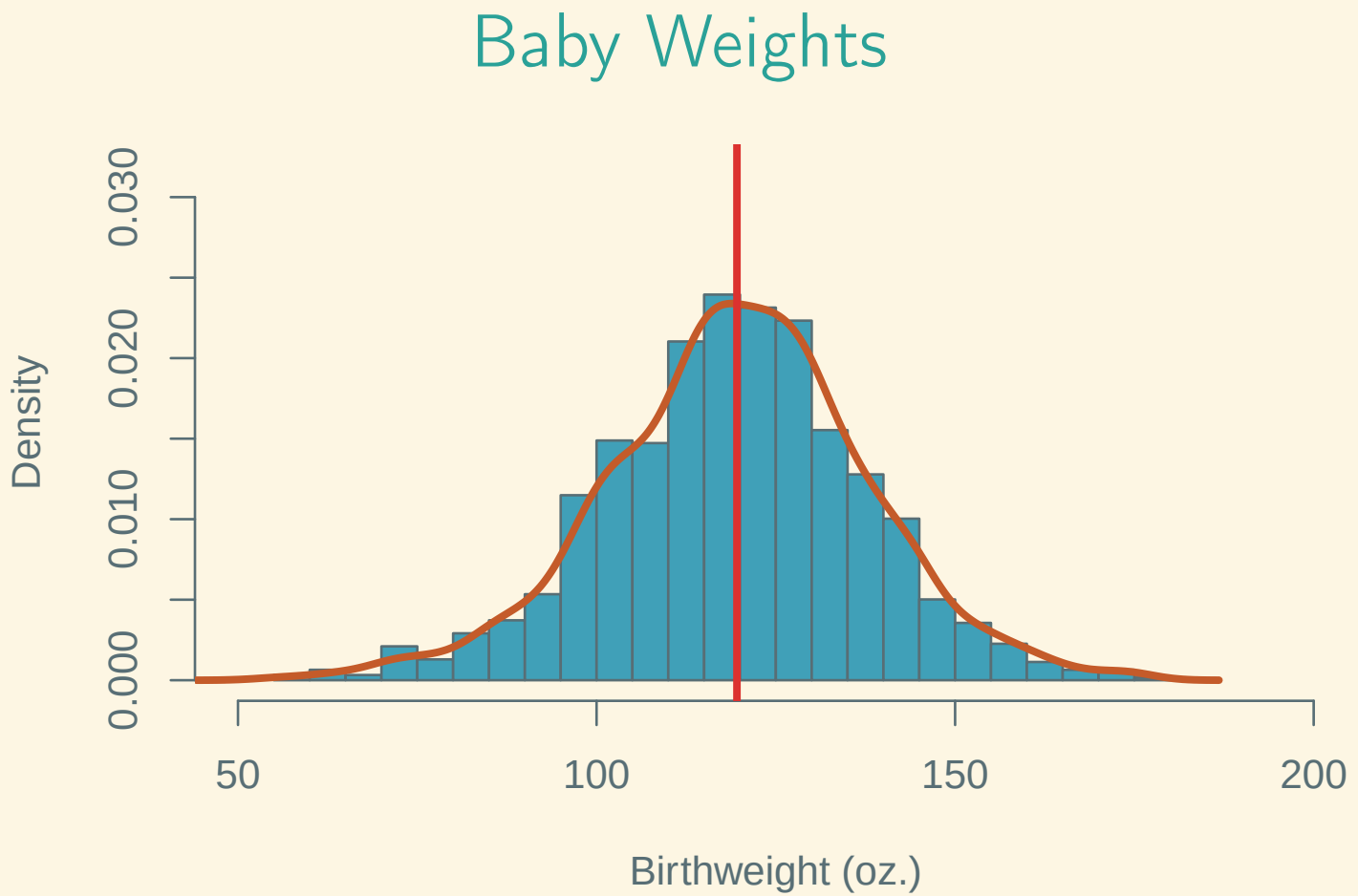


Figure: Distribution of Babies' Birth Weights (Nolan and Speed, 2000)

Household Incomes

Where's a typical value?

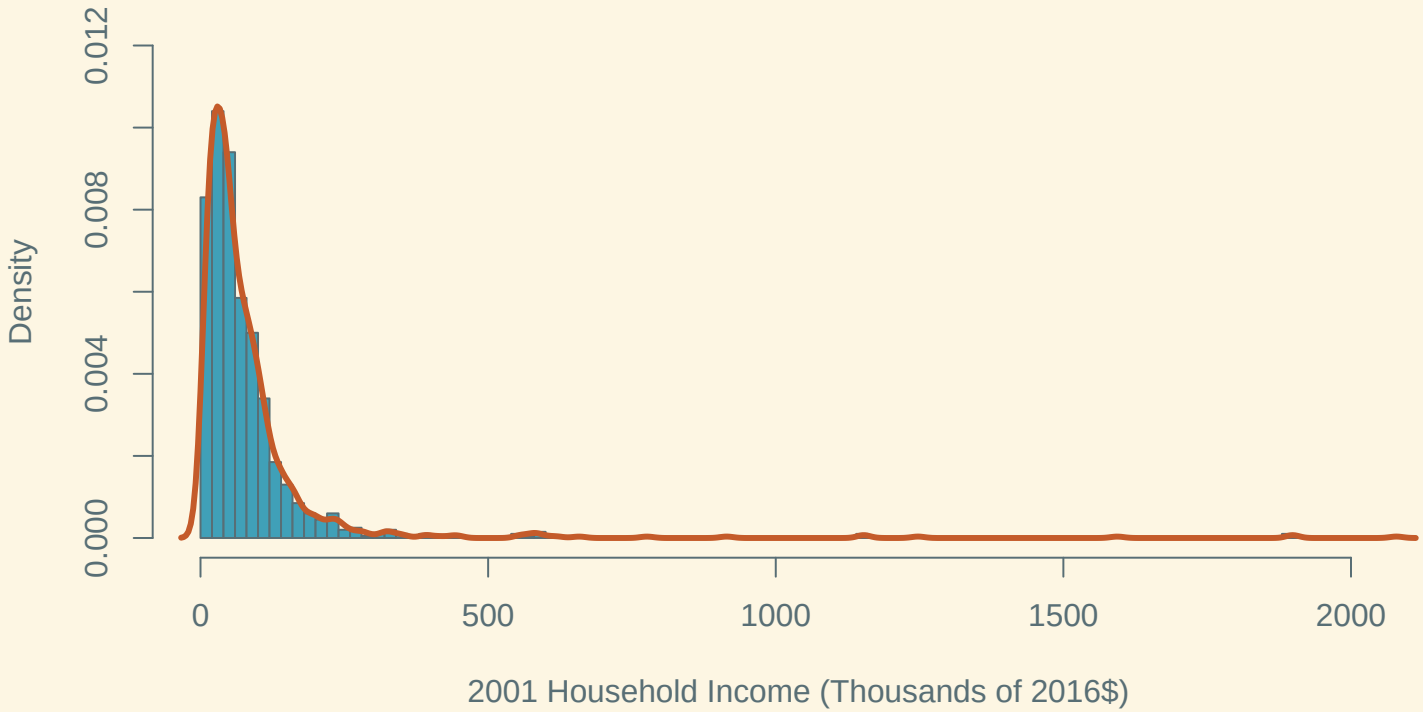


Figure: Distribution of Household Income (Source: 2001 Survey of Consumer Finances)

More than 70% of households are below average!

Household Incomes

Where's a typical value?

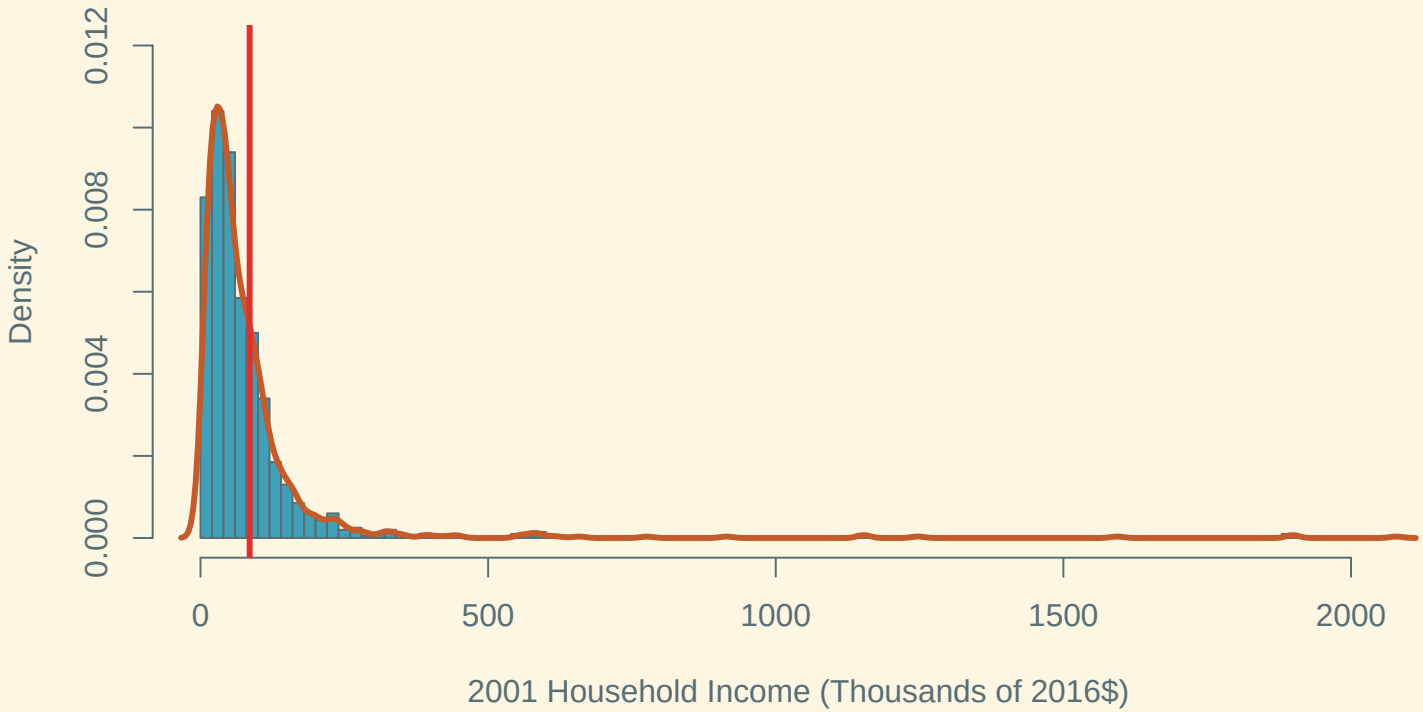


Figure: Distribution of Household Income (Source: 2001 Survey of Consumer Finances)

More than 70% of households are below average!

Household Incomes

How is this possible?

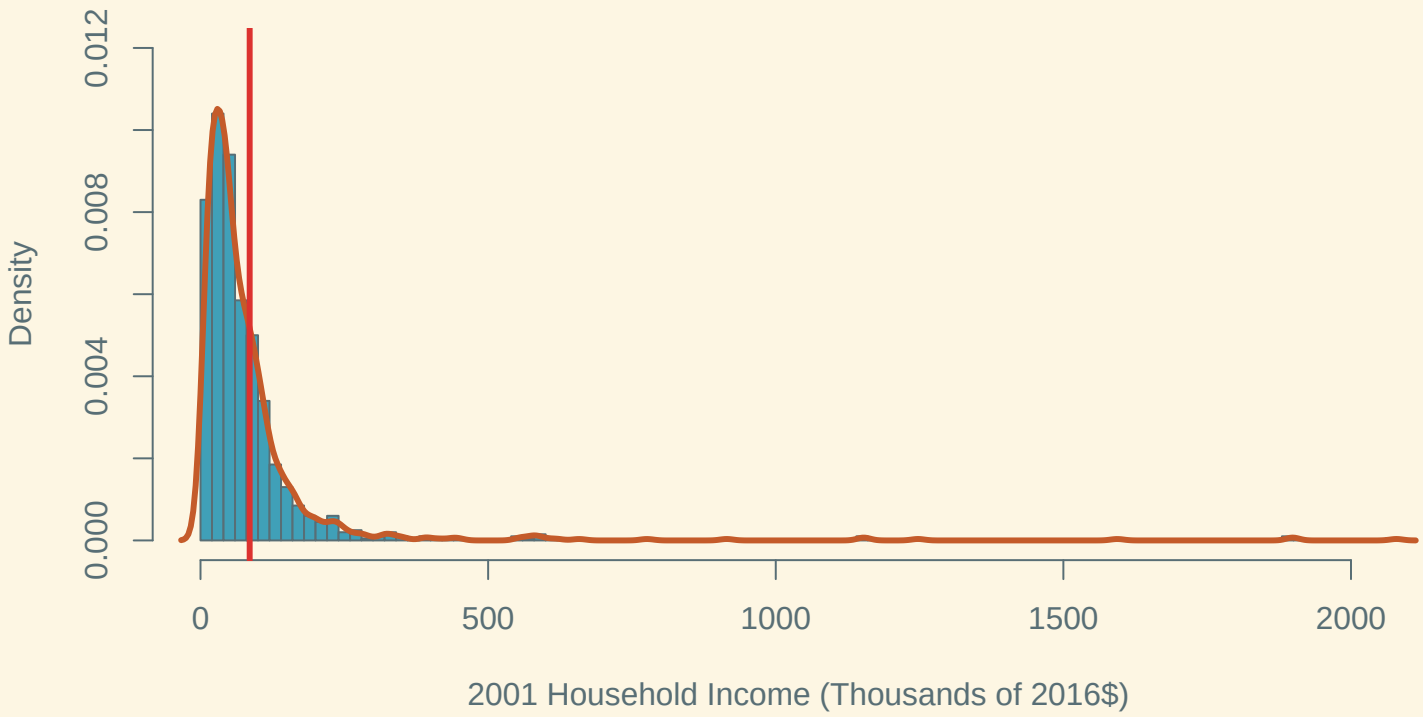
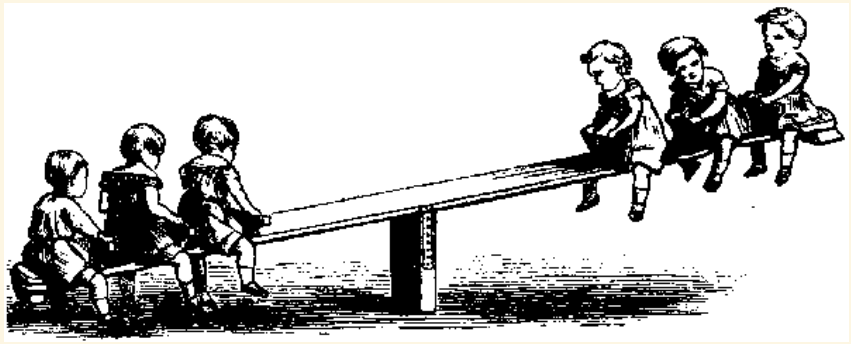


Figure: Distribution of Household Income (Source: 2001 Survey of Consumer Finances)

The Mean



Intuitively, the mean is the "balance point" of the data.

Mean

$$\bar{x} = (\sum_{i=1}^{n} x_i)/n$$

- x_i is the i^{th} observation
- n is the **sample size** (number of cases)

Thurston Howell, III walks into a bar...

Ten middle class friends are hanging out in a bar Their mean income is $36.7 K.

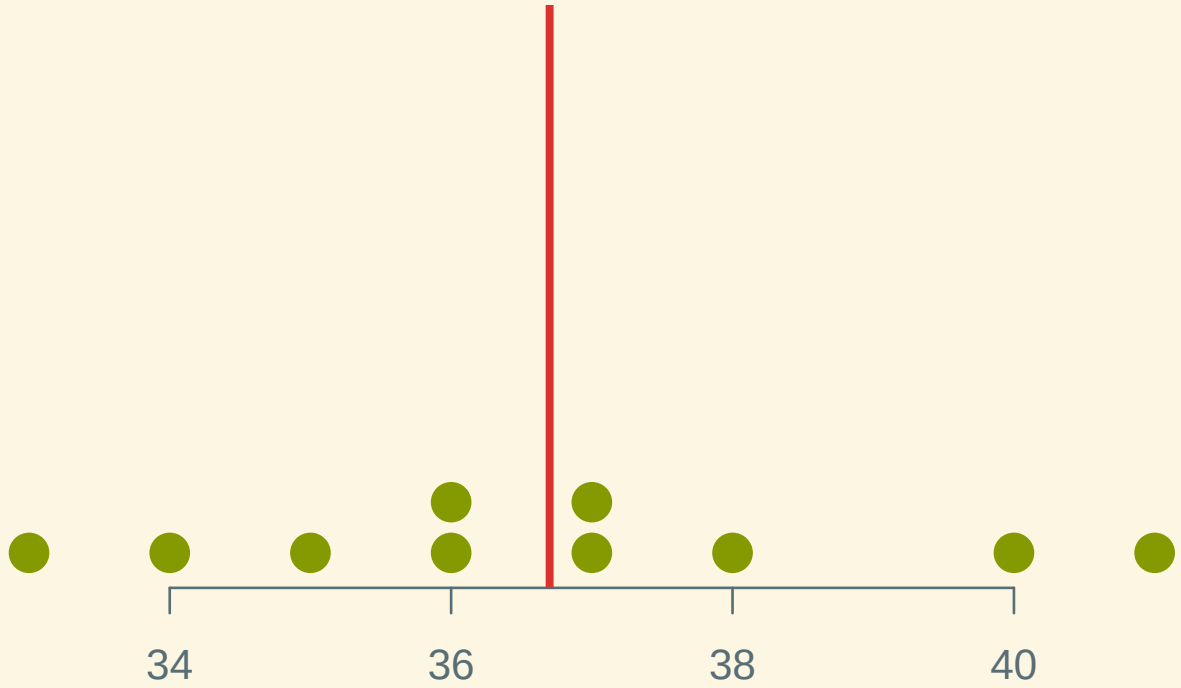


Figure: Incomes of Ten Barflies in Averageville, MO

Thurston Howell, III walks into a bar...

The highest earner walks out, and in walks Thurston Howell, III, the millionaire from Gilligan's Island. What happens to the mean income?

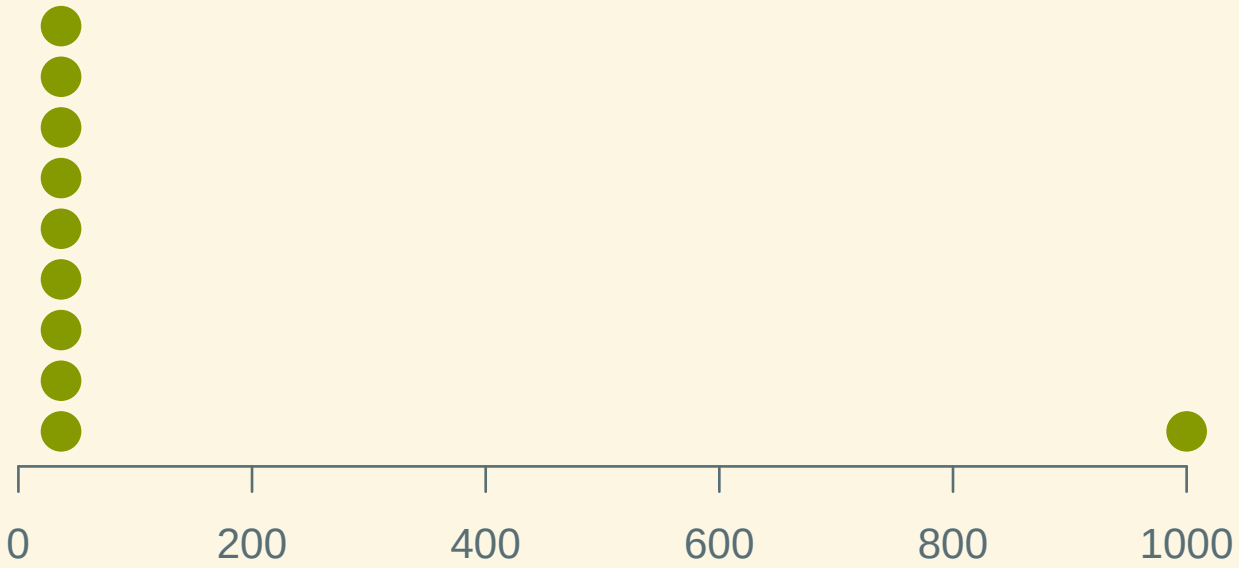


Figure: Incomes of Ten Barflies in Averageville, MO

Thurston Howell, III walks into a bar...

The highest earner walks out, and in walks Thurston Howell, III, the millionaire from Gilligan's Island. What happens to the mean income?

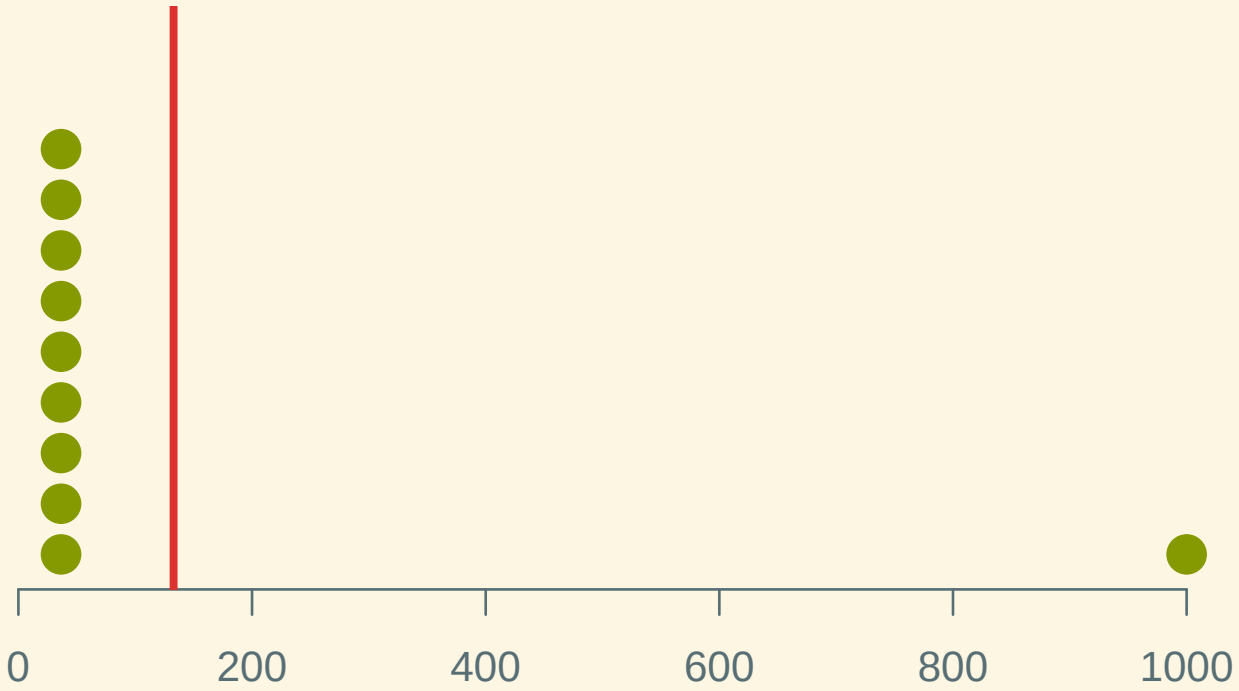


Figure: Incomes of Ten Barflies in Averageville, MO

Skewed Distributions

- The mean is strongly affected by skew and by **outliers**
- The mean is pulled toward the extreme values.
- In these cases, we generally prefer a measure of central tendency which is **resistant** to the influence of extreme values (also called **robust**).

The Median

Median

The **median** is the point that cuts the data in half: an equal amount of data lies above and below the median.

Thurston Howell, III walks into a bar...

- There's an even number of barflies, so the median is halfway between the fifth and sixth lowest income, in this case $36.5K

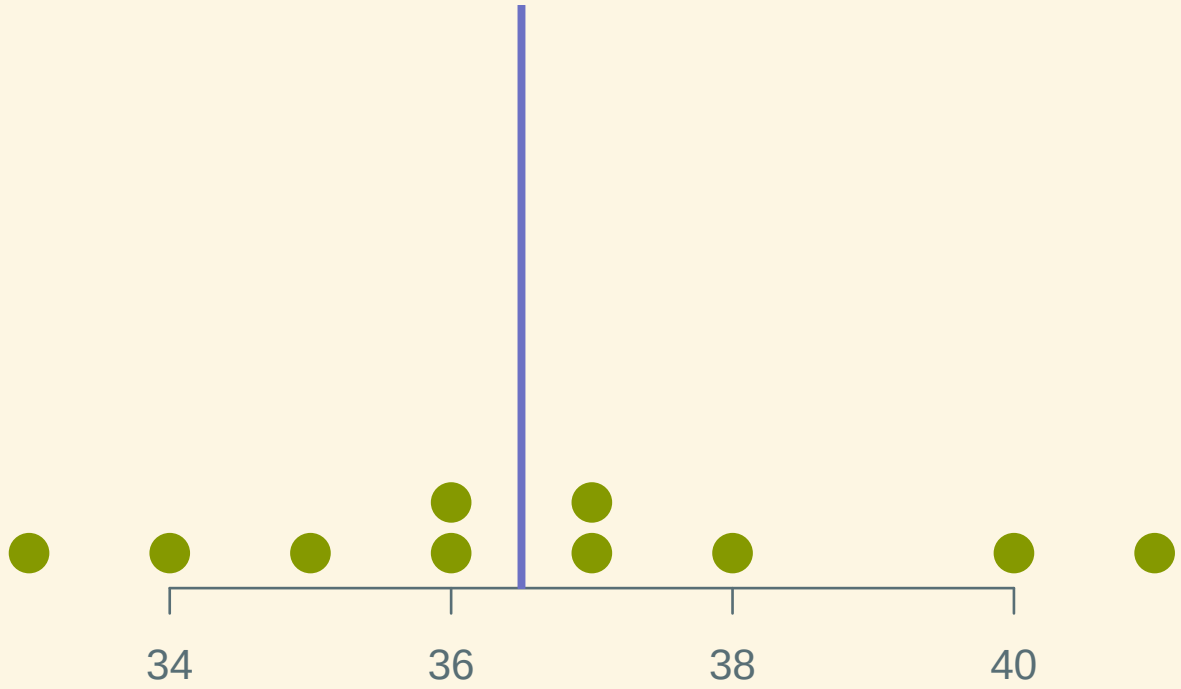


Figure: Incomes of Ten Barflies in Averageville, MO

Thurston Howell, III walks into a bar...

- When Thurston Howell, III walks in, what happens to the median?

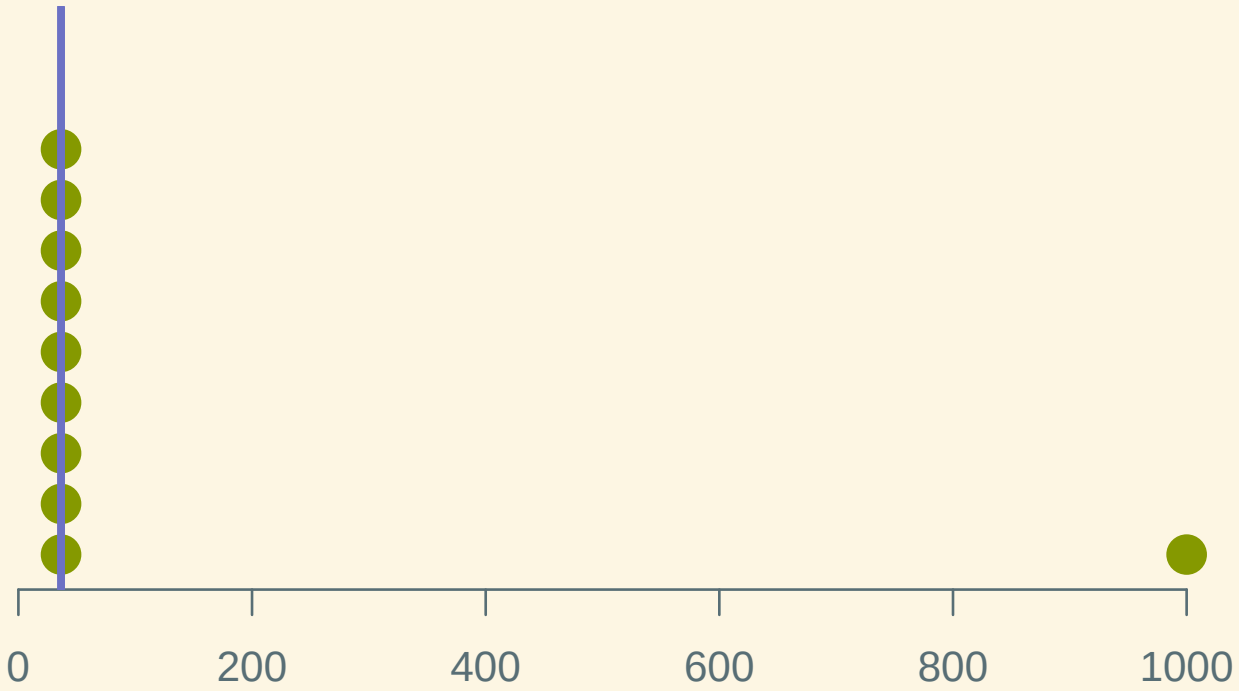


Figure: Incomes of Ten Barflies in Averageville, MO

The Median

The mean household income in 2001 (adjusted for inflation) was \$85.5K.

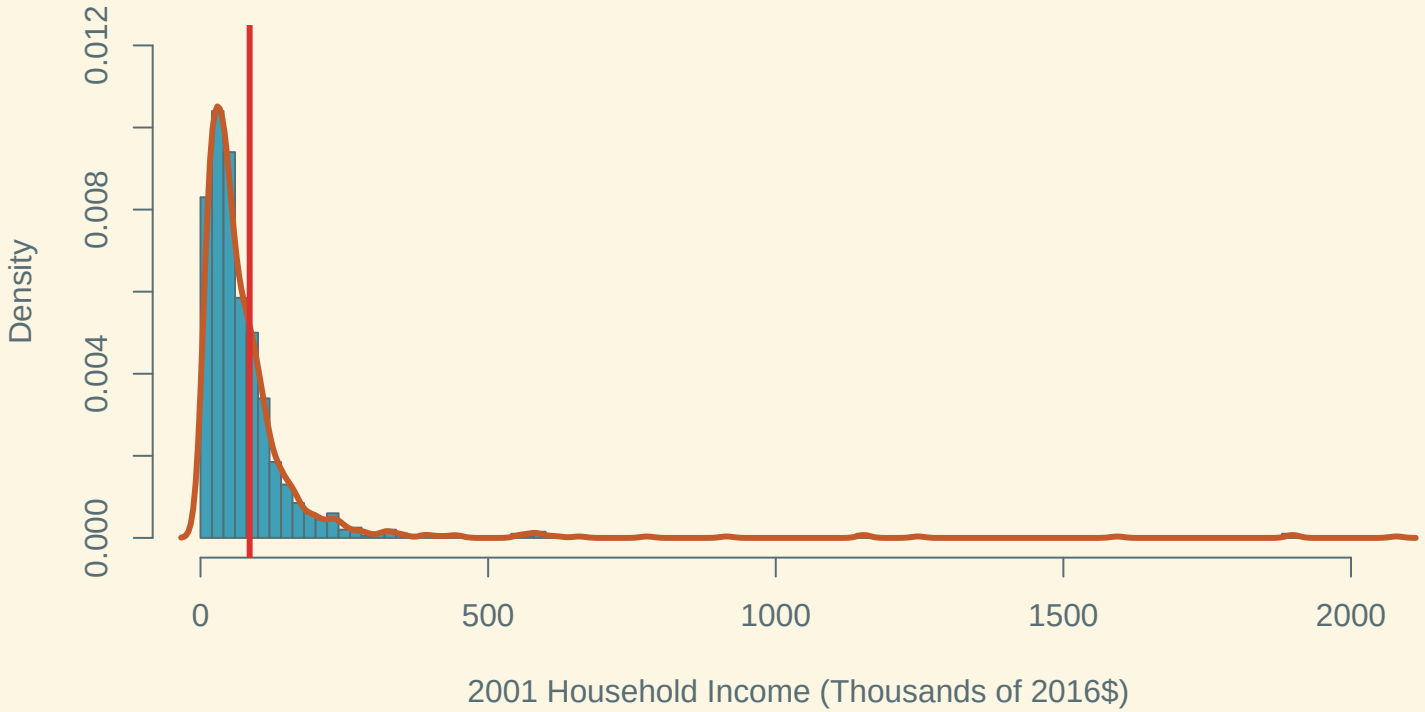


Figure: Distribution of 2001 Household Income

The Median

Where will the median be?

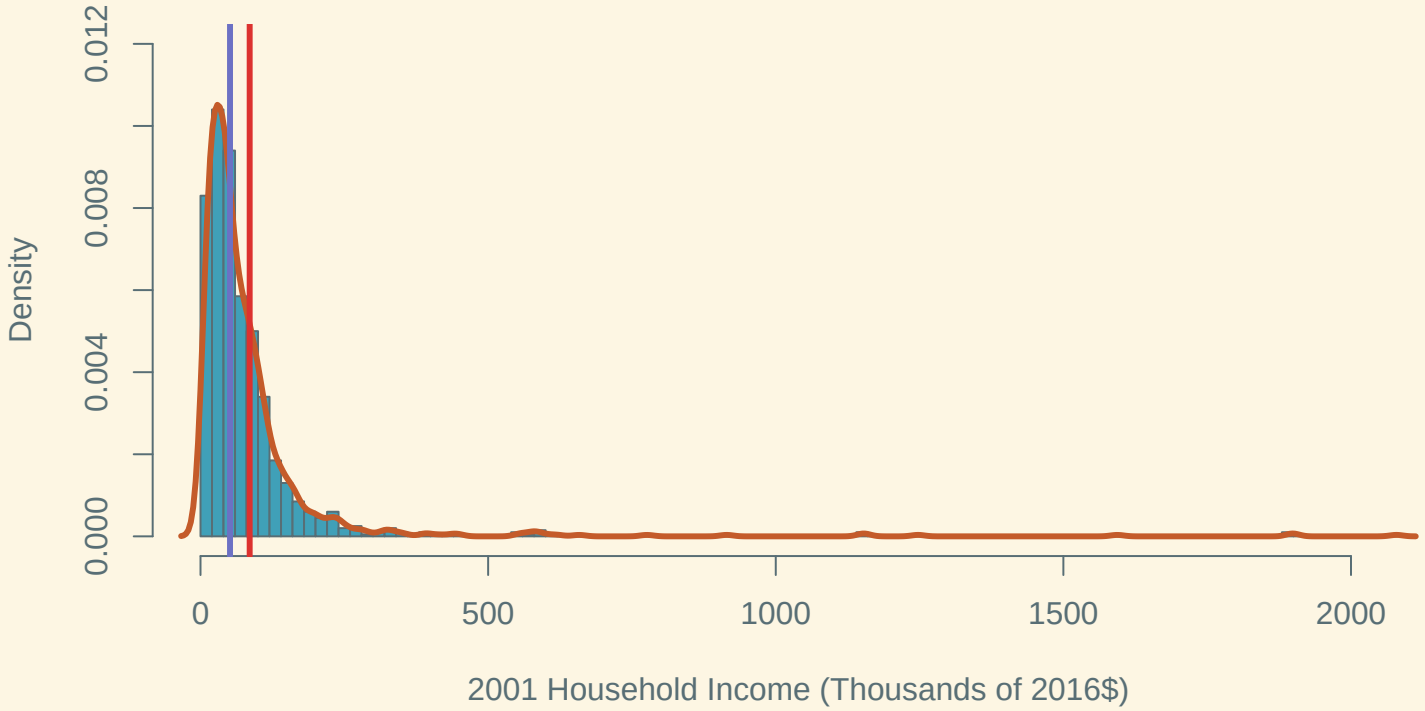


Figure: Distribution of 2001 Household Income

The Median

It is less affected by extreme values, so it is nearer to the bulk of households.

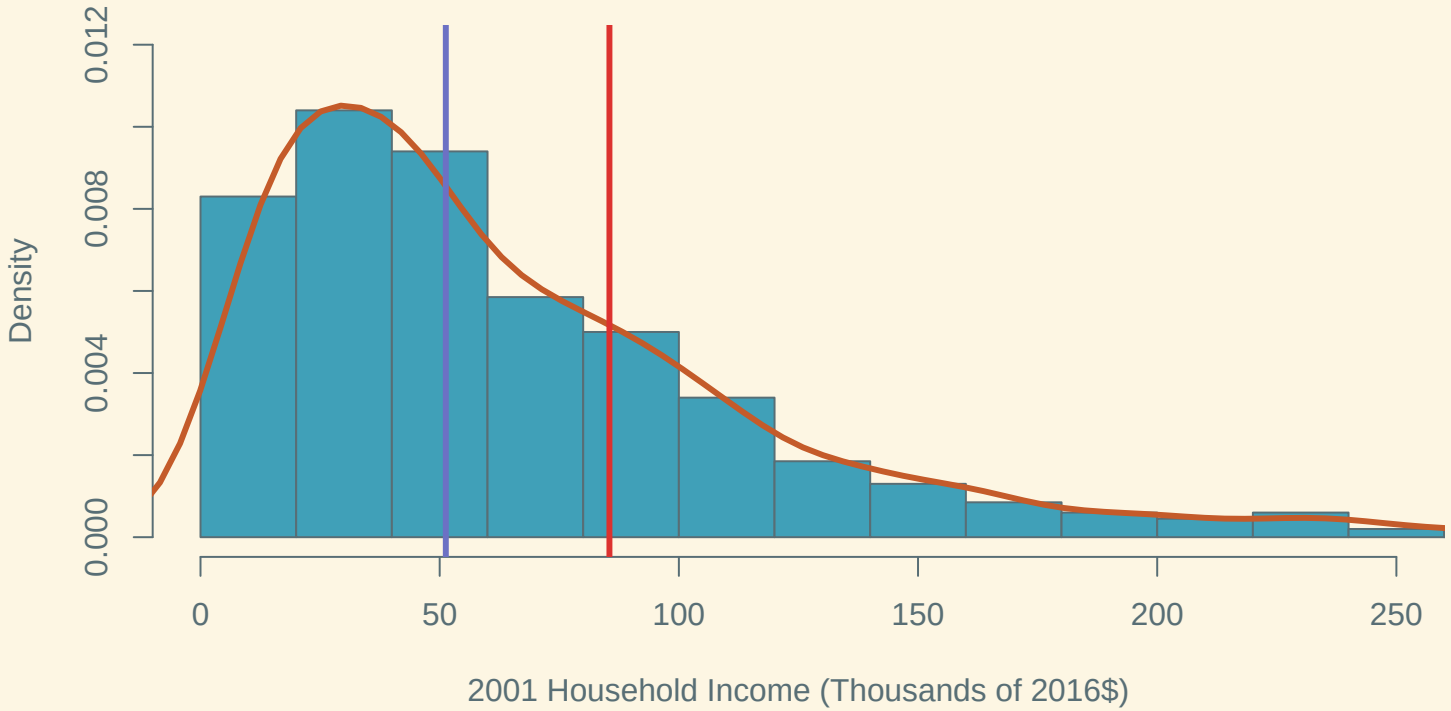


Figure: Distribution of 2001 Household Income

Median vs. Mean Income

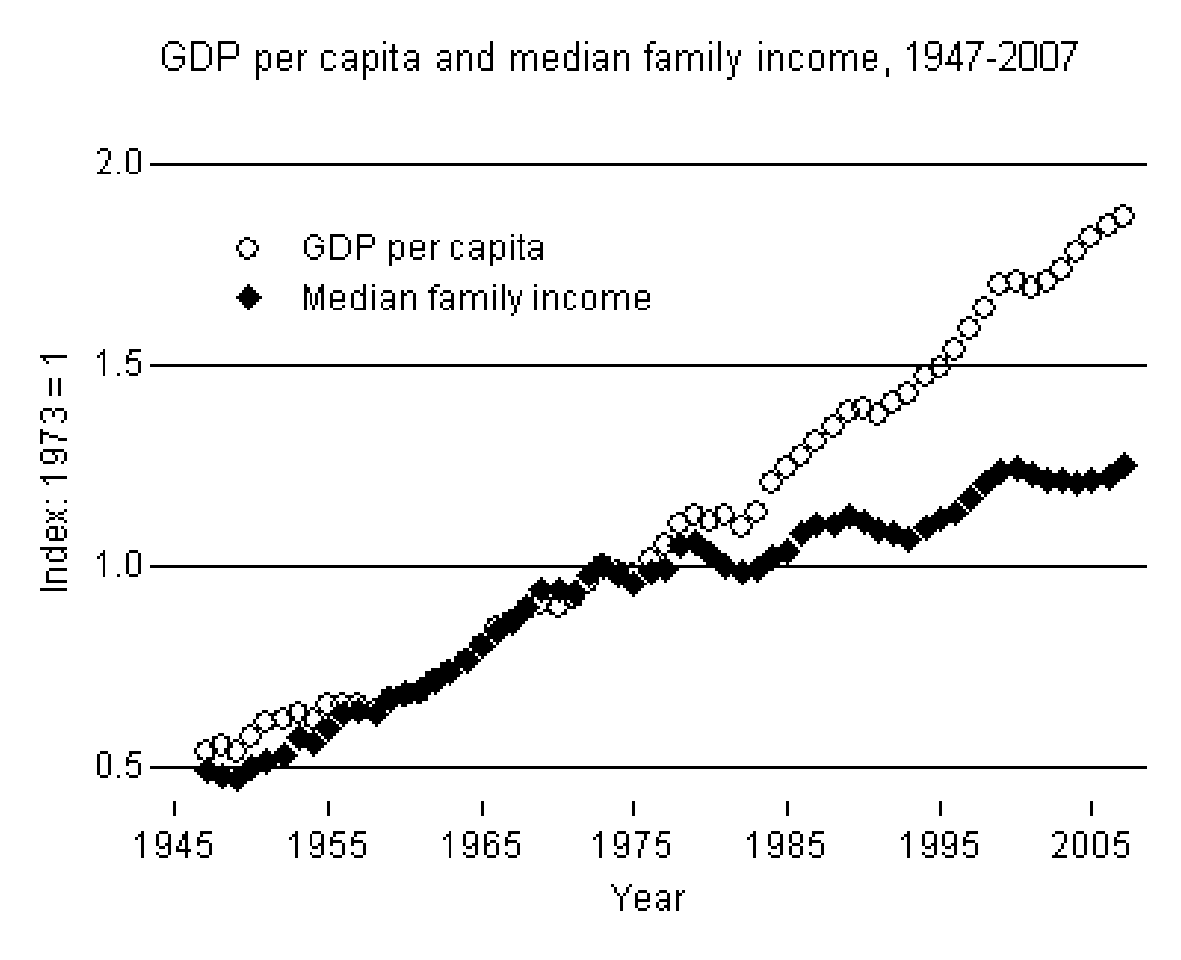


Figure: Median Household Income vs. Per Capita GDP (1947-2007).
Source: www.lanekenworthy.net

Median vs. Mean Income

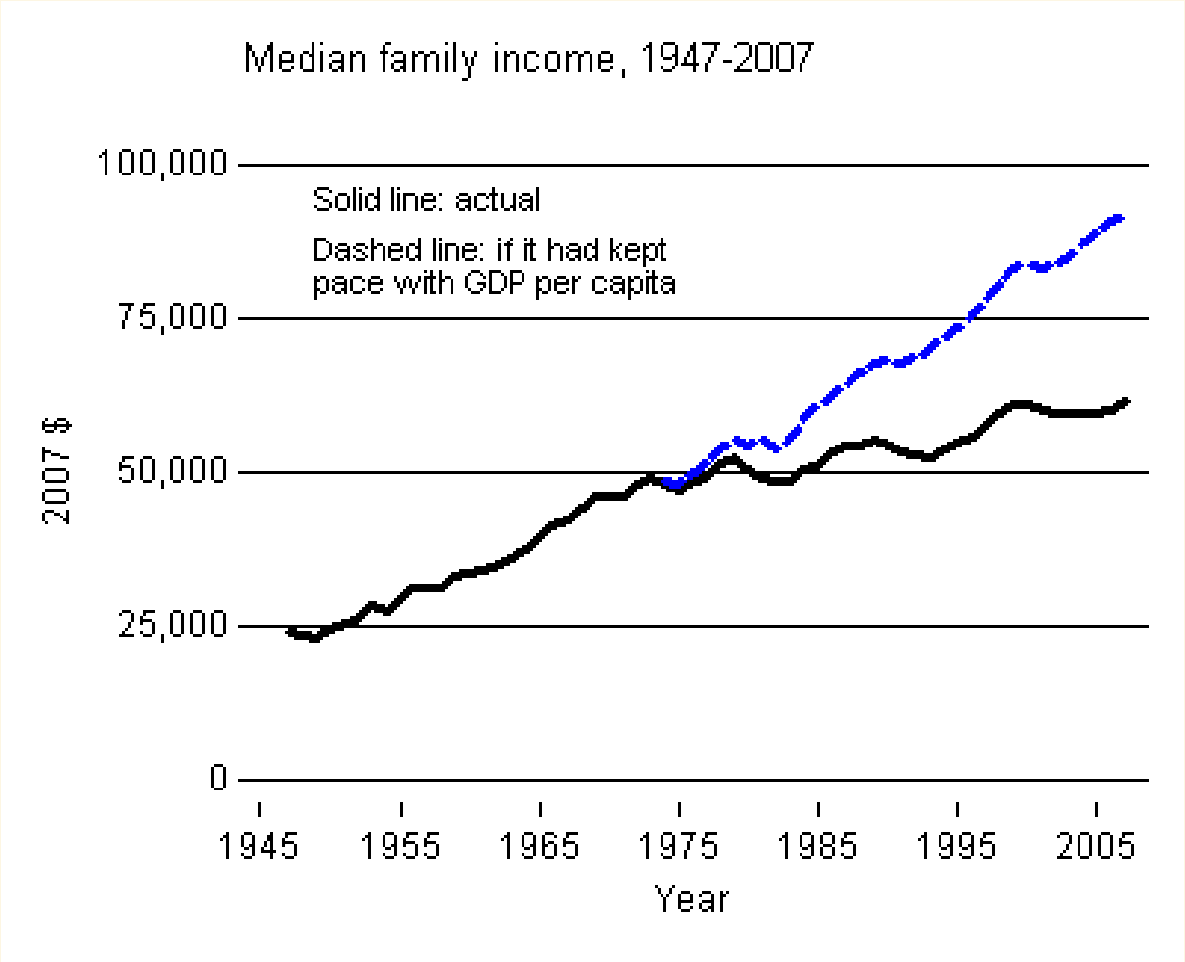


Figure: Median Household Income vs. Per Capita GDP (1947-2007)