# STAT/SOC/CSSS 221
# Statistical Concepts and Methods for the Social Sciences

## Summarizing Data

Christopher Adolph

Department of Political Science

*and*

Center for Statistics and the Social Sciences

University of Washington, Seattle

# Aside on mathematical notation

$\hat{\sigma}$ a "hat" on a variable indicates an estimate of an unknown quantity

# Aside on mathematical notation

| | |
|---|---|
| $\hat{\sigma}$ | a "hat" on a variable indicates an estimate of an unknown quantity |
| $x_{\text{label}}$ | sometimes, we note a way to describe a specific variable; we often used a subscripted label |

# Aside on mathematical notation

| | |
|---|---|
| $\hat{\sigma}$ | a "hat" on a variable indicates an estimate of an unknown quantity |
| $x_{\text{label}}$ | sometimes, we note a way to describe a specific variable; we often used a subscripted label |
| $x_i$ | for a set of variables $x_1$, $x_2$, ..., we can refer to an arbitrary member of this set using $x_i$ |

# Aside on mathematical notation

| | |
|---|---|
| $\hat{\sigma}$ | a "hat" on a variable indicates an estimate of an unknown quantity |
| $x_{\text{label}}$ | sometimes, we note a way to describe a specific variable; we often used a subscripted label |
| $x_i$ | for a set of variables $x_1$, $x_2$, ..., we can refer to an arbitrary member of this set using $x_i$ |
| $\sum_{i=1}^{n} x_i$ | the large capital sigma at left takes the sum over the set $x_1, x_2, \ldots, x_i, \ldots, x_n$ |

# Aside on mathematical notation

| | |
|---|---|
| $\hat{\sigma}$ | a "hat" on a variable indicates an estimate of an unknown quantity |
| $x_{\text{label}}$ | sometimes, we note a way to describe a specific variable; we often used a subscripted label |
| $x_i$ | for a set of variables $x_1$, $x_2$, ..., we can refer to an arbitrary member of this set using $x_i$ |
| $\sum_{i=1}^{n} x_i$ | the large capital sigma at left takes the sum over the set $x_1, x_2, \ldots, x_i, \ldots, x_n$ |
| E(x) | the *expected value* of $x$, where $x$ is typically a random variable |

# Aside on mathematical notation

| | |
|---|---|
| $\hat{\sigma}$ | a "hat" on a variable indicates an estimate of an unknown quantity |
| $x_{\text{label}}$ | sometimes, we note a way to describe a specific variable; we often used a subscripted label |
| $x_i$ | for a set of variables $x_1$, $x_2$, ..., we can refer to an arbitrary member of this set using $x_i$ |
| $\sum_{i=1}^{n} x_i$ | the large capital sigma at left takes the sum over the set $x_1, x_2, \ldots, x_i, \ldots, x_n$ |
| E(x) | the *expected value* of $x$, where $x$ is typically a random variable |
| $\text{function}(x)$ | we can make a new math function by naming it, and putting the "inputs" to it in parentheses |

# Outline

Central Tendency of a Random Variable

Dispersion of a Random Variable

Exploring Data with Graphics (part 1)

# Running example: Household wealth in 2007

We take the following data from the 2007 Survey of Consumer Finances:

**Net Household Wealth** The sum of financial and non-financial assets (e.g., vehicle and home equity), minus debt, in thousands of dollars

**Education** The education of the head of household, coded as less than high school, high school, some college, and college

**Age** The age of the head of household, in years

**Race** The self-identified race of the head of household: non-Hispanic white, black, Hispanic, Asian, or other

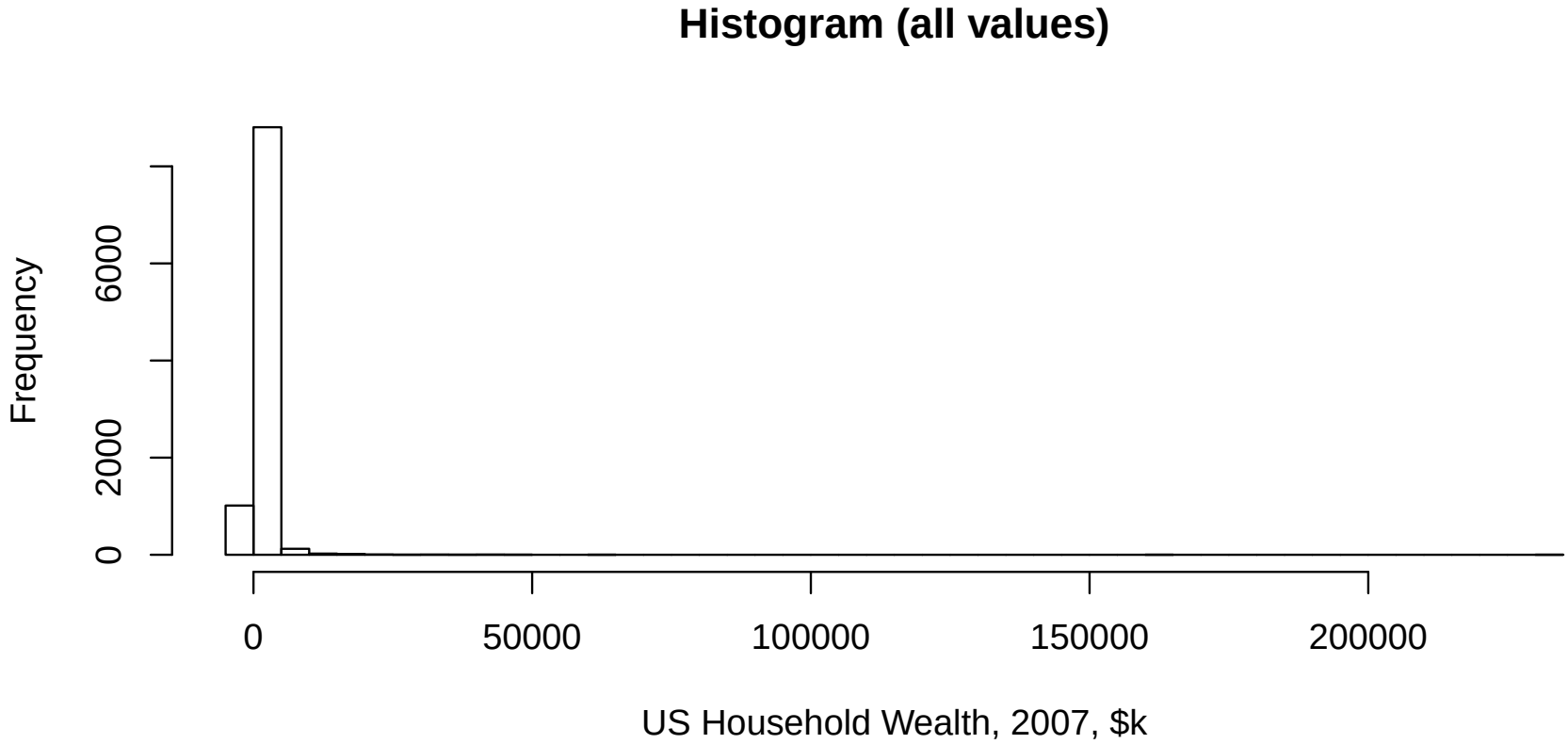
# Running example: Household wealth in 2007

The first few samples from this data are:

| wealth | educ | age | race |
|---|---|---|---|
| −0.40 | Less than HS | 48 | White |
| 689.00 | Less than HS | 67 | White |
| 197.40 | College | 44 | White |
| −9.25 | High School | 51 | White |
| 19.02 | High School | 49 | White |
| −3.66 | College | 36 | White |

Our primary interest is in *wealth* in thousands of dollars

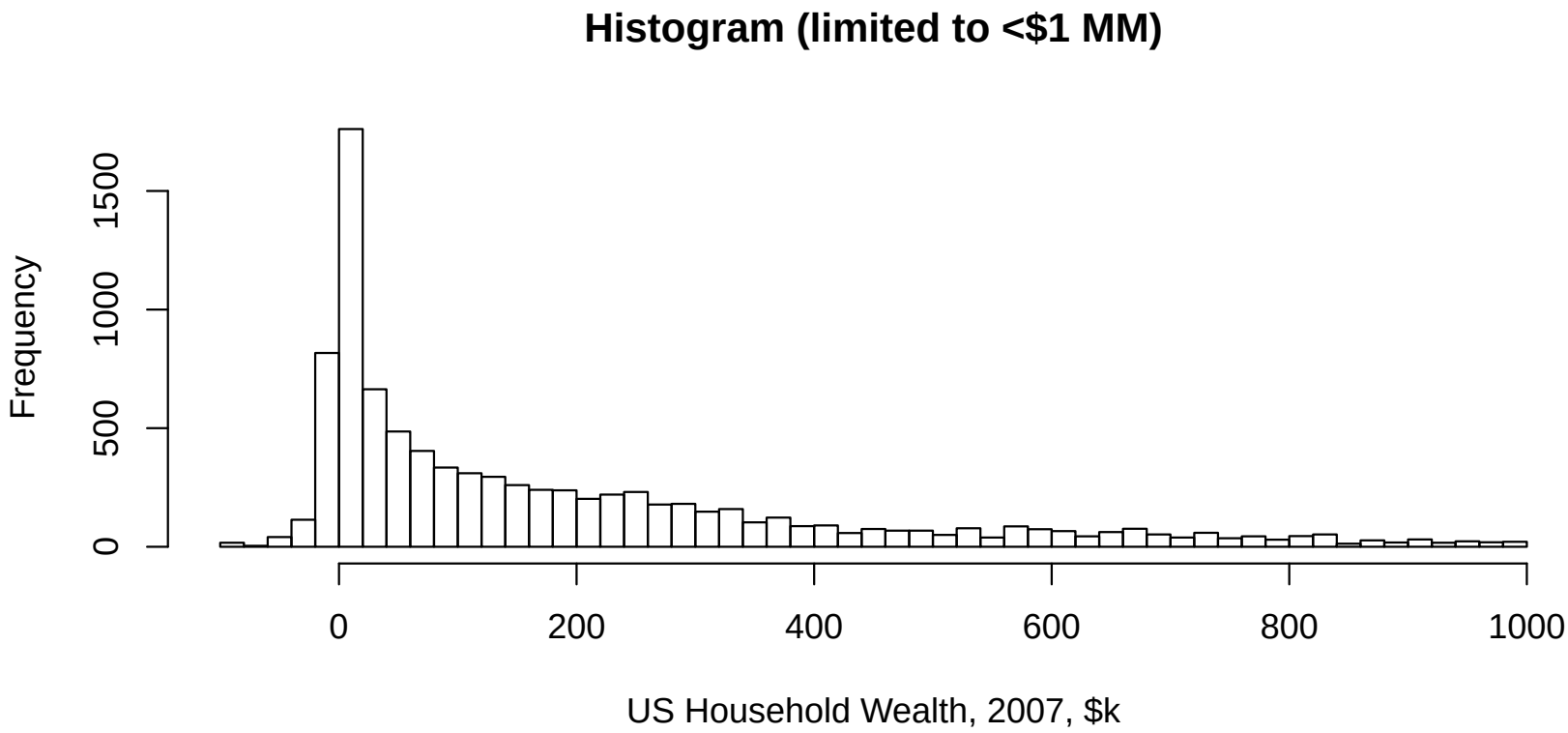
We have 10000 samples:
Can we get an overview without poring over a hundred slides of numbers?




Make a histogram!

One problem:
the wealth distribution has mostly small values and a few very large ones

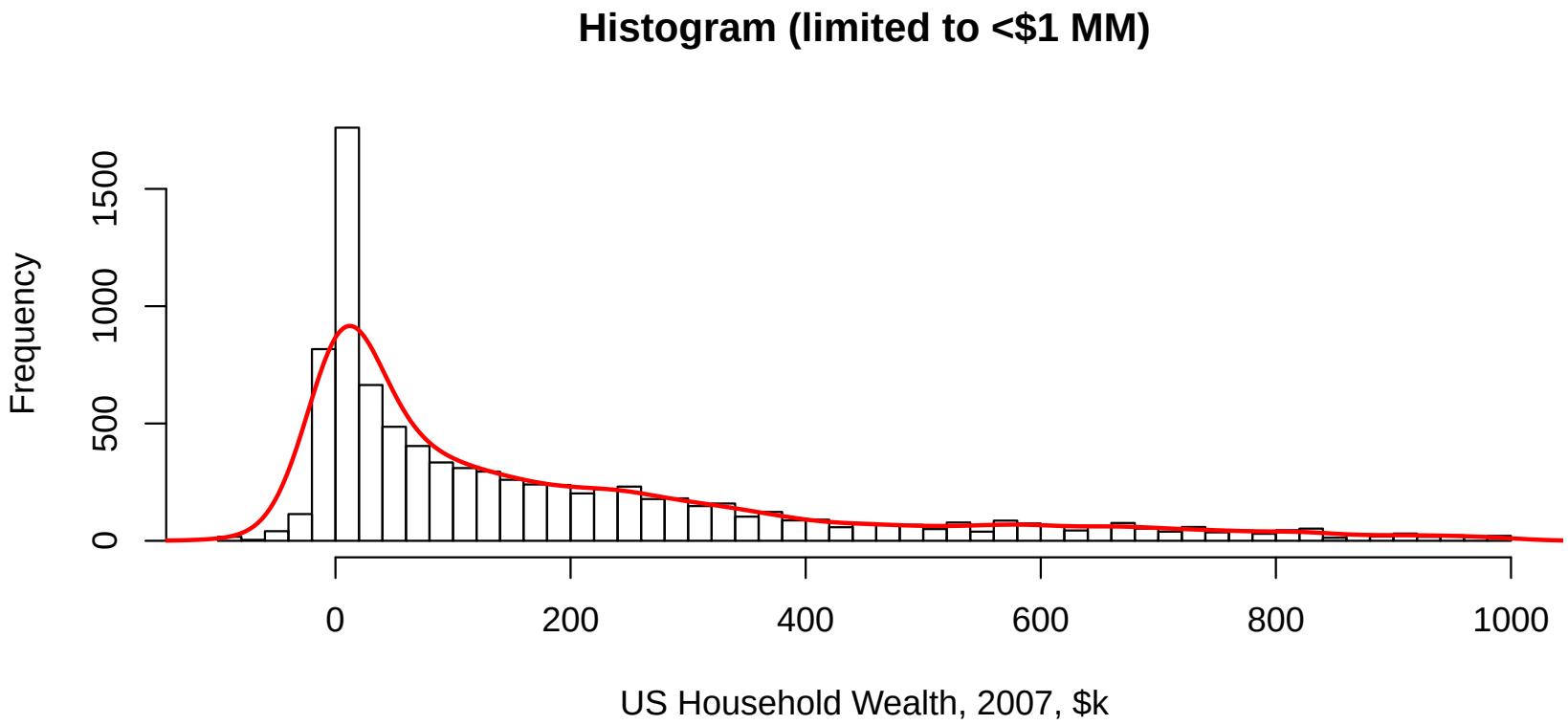
(Bill Gates is skewing the scale and pushing the rest of us into the left bin)




Let's restrict the picture to households with less than one million in wealth

Note that the choice of "bins" for the histogram is up to us

Our program may default to bins that are too wide or too narrow




Finally, we can replace the whole histogram with a density plot

The red line was produced by a kernel density estimator (technical)

Don’t need to know how to make, just that it can help clean up the graph

# Random variables review

| wealth | educ | age | race |
|---|---|---|---|
| −0.40 | Less than HS | 48 | White |
| 689.00 | Less than HS | 67 | White |
| 197.40 | College | 44 | White |
| −9.25 | High School | 51 | White |
| 19.02 | High School | 49 | White |
| −3.66 | College | 36 | White |

Quick review: what are the levels of measurement here?

## Review from last time

Recall the three different ways to divide random variables into categories:

- Discrete vs. Continuous
- Nominal vs. Ordered
- Additive vs Ratio Intervals

But what exactly is a random variable?

A random variable is a variable with some probability of taking on each of several values

# Summary statistics

There are two broad ways to summarize how a sample from a random variable behaves:

1. Identify the central tendency of a variable
2. Measure the dispersion around that center

# Defining measures of central tendency

Consider a random variable $x$ with $n_x = 15$ observations:

3 8 9 4 3 8 1 8 7 8 1 5 3 10 8

It will help to sort this variable from smallest to largest.

Define $x_{\text{sorted}} = \text{sort}(x)$:

1 1 3 3 3 4 5 7 8 8 8 8 8 9 10

How can we summarize the central tendency of this variable?

Three options: the **mode**, **median**, and **mean**

# The Mode of a Random Variable

The mode of a variable is its most common value:

1 1 3 3 3 4 5 7 8 8 8 8 8 9 10

In this case, 8 is the most frequently occurring value.

The mode is clearly defined for **any discrete** variable.
Doesn't matter if the variable is ordered or nominal

The mode is not clearly defined for continuous variables

# The Mode of a Random Variable

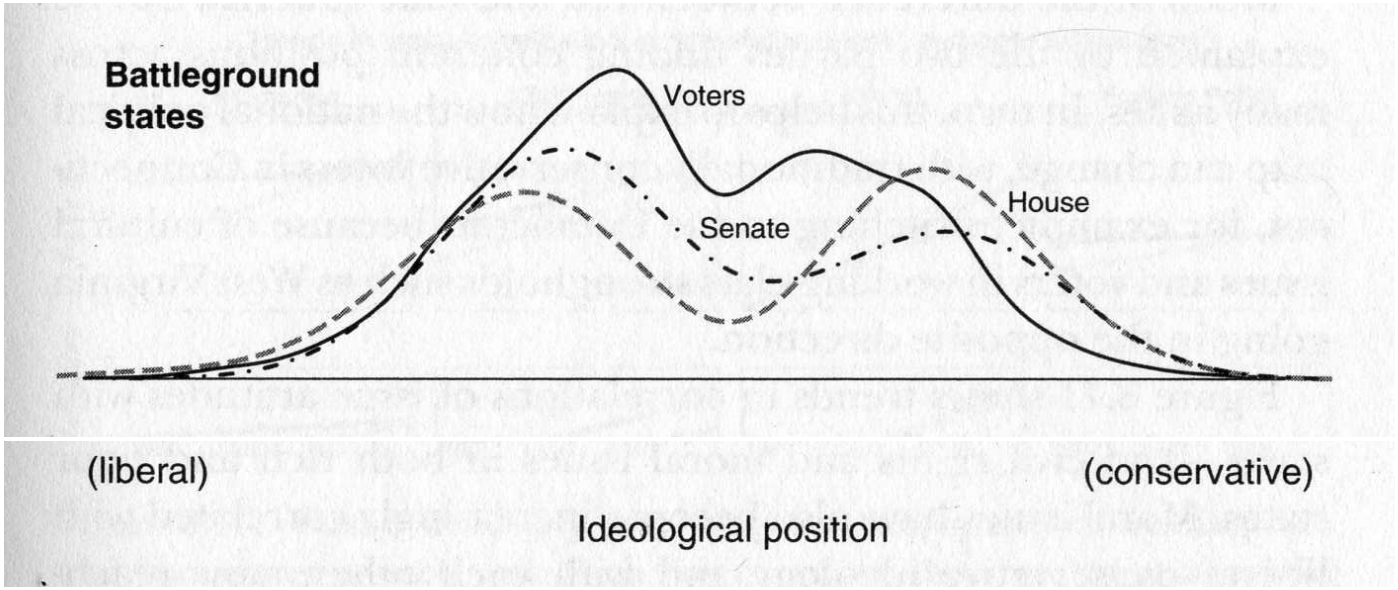
The probability that a continuous variable takes on any specific value is 0

There are just too many possible values for any one to repeat

So there won't be a clear mode in the literal sense

However, social scientists often speak of the mode of a continuous variable with reference to its density

# The Mode of a Random Variable



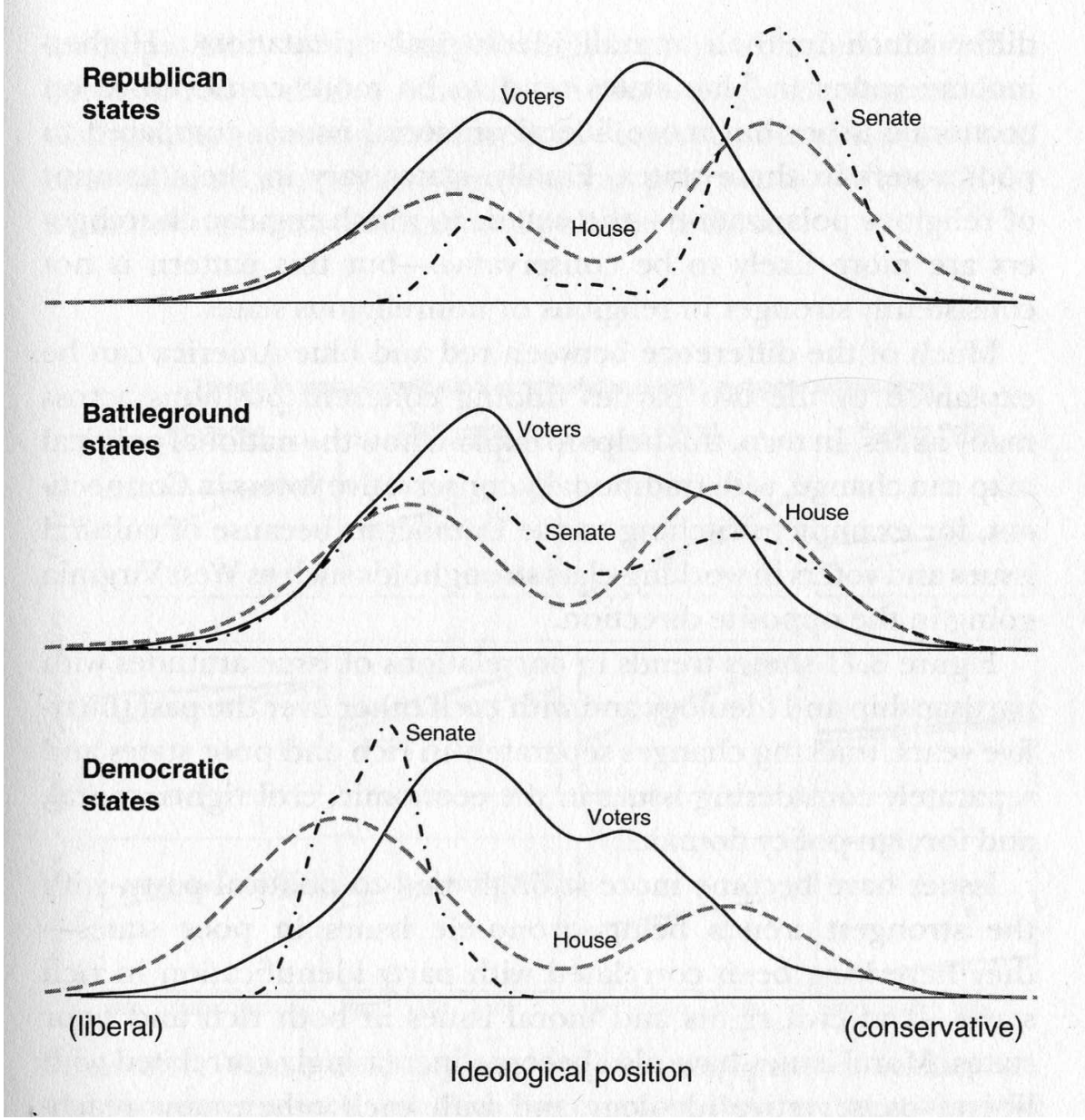

Each maximum on this density graph can be though of as a “mode” (Graph from Gelman, *Red State, Blue State*)

If more than one maximum appears, we call the variable “multi-modal” (the above distributions all appear “bimodal”)

If only one maximum appears, we call the variable “unimodal”

Any maximum on the density plot that “stands-out” visually is a mode, whether or not it is the global maximum

# The Mode of a Random Variable



# The Median of a Random Variable

The median of a variable is its middle value:

1 1 3 3 3 4 5 7 8 8 8 8 8 9 10

If $n_x$ is odd, the median is the middle value of the sorted list.

If $n_x$ is even, we can just average the middle pair of values.

Here, of course, the median is 7.

# The Median of a Random Variable

The median has two nice properties:

1. It is defined for any ordered random variable
(either continuous or discrete)

2. It is *robust* or *resistant*, in the sense that adding a big outlier to the data does not move the median "too much"

   For example, suppose we in recording our random variable,
   we accidentally mistyped 10 as 100.

   The mean would double, but the median would still be 7

# The Mean of a Random Variable

The mean of a random variable is. . .

# The Mean of a Random Variable

The mean of a random variable is. . . what exactly?

1 1 3 3 3 4 5 7 8 8 8 8 8 9 10

# The Mean of a Random Variable

The mean of a random variable is. . . what exactly?

1 1 3 3 3 4 5 7 8 8 8 8 8 9 10

The mean is the best known measure of central tendency

We know how to calculate it:

$$\text{mean}(x) = \bar{x} = \frac{1}{n_x} \sum_{i=1}^{n_x} x_i$$

Here, the mean is $5.7\bar{3}$

But what is it in words?
(ie, the mode is the “most common value”, the median is “the middle value”, and the mean is. . . )

# Understanding measures of central tendency: Street Fair Example

At the Gauss Street fair,
you are offered the chance to win a prize for guessing a man's weight

But the rules of the contest are strict and unusual

- The man is behind a screen, so you can't judge based on appearance
- He has been drawn at random from the US population
- You must guess his *exact* weight, to the pound, or you win nothing!

What should you guess?

# Understanding measures of central tendency: Street Fair Example

At the Gauss Street fair,
you are offered the chance to win a prize for guessing a man's weight

But the rules of the contest are strict and unusual

- The man is behind a screen, so you can't judge based on appearance
- He has been drawn at random from the US population
- You must guess his *exact* weight, to the pound, or you win nothing!

What should you guess?

The *mode* of US males' weight (rounded to the pound)

This will be more often right than any other guess. (But only right how often?)

# Understanding measures of central tendency: Street Fair Example

Further down Gauss Street,
you find another unusual weight-guessing contest.

This time:

- There is one man behind a screen
- He has been drawn at random from the US population
- You must guess the man's weight to the nearest fraction of a pound. The closest guess wins.

What should you guess?

# Understanding measures of central tendency: Street Fair Example

Further down Gauss Street,
you find another unusual weight-guessing contest.

This time:

- There is one man behind a screen
- He has been drawn at random from the US population
- You must guess the man's weight to the nearest fraction of a pound. The closest guess wins.

What should you guess?

The *median* of US males' weight. For most randomly selected men, the median will be close to the right answer

# Understanding measures of central tendency: Street Fair Example

In unimodal continuous distributions,
the median tends to be close to as much of the data as possible

For most cases, the mean will be higher than the median because of the existence of a small handful of very obese individuals in the US population (outliers)

That is, if the average weight is 175, and there are some individuals weighing 375, there simply can't be any -25 pounders to balance them out in the mean.

Another way to put this: the distribution of weight is right-skewed

# Understanding measures of central tendency: Street Fair Example

At the last booth on Gauss Street, still another vendor seeks your hard-earned money in a weight-guessing contest

This time:

- There are 30 men behind a screen
- Each has been drawn at random from the US population
- The booth has a fancy scale to weighh this group all at once.
  You must guess their combined weight, knowing only how many men are in the group
- Everyone who makes a guess gets a prize,
  but the prize is bigger the closer you get to the true total weight

What should you guess?

# Understanding measures of central tendency: Street Fair Example

At the last booth on Gauss Street, still another vendor seeks your hard-earned money in a weight-guessing contest

This time:

- There are 30 men behind a screen
- Each has been drawn at random from the US population
- The booth has a fancy scale to weighh this group all at once.
  You must guess their combined weight, knowing only how many men are in the group
- Everyone who makes a guess gets a prize,
  but the prize is bigger the closer you get to the true total weight

What should you guess?

The *mean* of US males' weight ($\times n_x$, of course)

# Understanding measures of central tendency: Street Fair Example

Some properties of the mean:

- The mean minimizes the sum of the (squared) distances between the population and itself

# Understanding measures of central tendency: Street Fair Example

Some properties of the mean:

- The mean minimizes the sum of the (squared) distances between the population and itself
- The mean minimizes the (squared) error of prediction of a random draw from the population

# Understanding measures of central tendency: Street Fair Example

Some properties of the mean:

- The mean minimizes the sum of the (squared) distances between the population and itself
- The mean minimizes the (squared) error of prediction of a random draw from the population
- The mean is the expected value of x. That is

$$\mathrm{E}(x) = \mathrm{mean}(x) = \bar{x} = \frac{1}{n_x} \sum_{i=1}^{n_x} x_i$$

# When can we use the mean?

Can the mean be computed for all kinds of random variables?

## When can we use the mean?

Can the mean be computed for all kinds of random variables?

No, just additive or ratio level variables

But remember we can convert nominal variables to a series of binary ones, and binary variables are ratio level

So we can convert our education variable to four binary variables, and calculate their means:

| | |
|---|---|
| Less than High School | 0.14 |
| High School | 0.33 |
| Some College | 0.19 |
| College | 0.34 |

For binary variables,
the mean is the only appealing measure of central tendency (why?)

# Robustness

A robust statistic is one that is relatively insensitive to extreme values

Extreme values are often misleading:

- They may be simple measurement error; e.g., the coder added a digit
- They may result from unusual processes not similar to the middle of the data; e.g., the origins of internet billionaires' wealth

Non-robust statistics like the mean may reflect the middle of our data poorly compared to robust measures like the median

However, if there are no weird outliers, non-robust measures will be more efficient in using all the available information

# Grading example

Most teachers compute course grades by taking the mean of a set of assignment grades.

Why use the mean, and not the median or mode?

When would the mean be the “best” way to compute a course grade?

When would the median be better?

Would we ever want the mode?

# Measuring dispersion

Measures of central tendency leave out a lot compared to histograms

If summarizing a histogram in one number, use a central tendency

But if you had time for a two number summary? What would you want next?

Some sense of *dispersion*, or how far away the data tend to be from the center

# Measuring dispersion

Compared to central tendency, dispersion is easy to overlook, but crucial:

1. Tells us how well the center predicts specific values
2. Tells us how noisy the variable is
3. Tells us how much the world differs from observation to observation

Because social science is about understanding variation,
dispersion should be our bread and butter

Three measures of dispersion: range, quantiles, and variance

# The Range of a Random Variable

The range of a variable is the set of its minimum and maximum values:

1 1 3 3 3 4 5 7 8 8 8 8 8 9 10

The range of this variable is [1,10]

The range is the simplest measure of dispersion, but the least useful

Very non-robust: Shifted by the most extreme outliers

The range is defined for any ordered random variable, discrete or continuous

# The Quantiles of a Random Variable

The quantiles of a random variable are taken at specific intervals of the sorted variable

1 1 3 3 3 4 5 7 8 8 8 8 8 9 10

For example, we might want the quantiles enclosing the middle half of the data

Those are the 25th and 75th percentiles, highlighted above

Note that the median is the 50th percentile, so the median is a quantile as well

Quantiles are defined for any ordered random variable

## The Quantiles of a Random Variable

We could calculate any number of quantiles.

For the wealth data, we have the following quantiles ($k):

| | |
|---|---|
| 0.1% | -106 |
| 1.% | -31 |
| 2.5% | -15 |
| 5.% | -5 |
| 25.% | 13 |
| 50.% | 115 |
| 75.% | 365 |
| 95.% | 1908 |
| 97.5% | 3597 |
| 99.% | 8409 |
| 99.9% | 25637 |

Quantiles are great: easy to understand, can summarize dispersion to arbitrary specificity, and are robust

# The Variance of a Random Variable

The variance of a variable is the square of the standard deviation:

1 1 3 3 3 4 5 7 8 8 8 8 8 9 10

Here, the variance is 9.0, and the standard deviation is $\approx 3.01$

But what is the standard deviation?

- how much a random draw from this random variable would differ from the mean on average
- the expected error when we predict this variable using its expected value

# Theoretical Variance of a Random Variable

The standard deviation has a simple meaning: “how much we miss by”

Mathematically, this take a bit of work:

$$\text{var}(x) = \sigma^2 \quad = \quad \text{E}\left((X - \text{E}(x))^2\right)$$

# Theoretical Variance of a Random Variable

The standard deviation has a simple meaning: “how much we miss by”

Mathematically, this take a bit of work:

$$\begin{aligned} \text{var}(x) = \sigma^2 &= \text{E}\left((X - \text{E}(x))^2\right) \\ \text{sd}(x) = \sigma &= \sqrt{\text{E}\left((X - \text{E}(x))^2\right)} \end{aligned}$$

In words, the variance is how much you expect a specific observation to differ from the mean of the variable, squared

We take the square root to get how much you expect a specific observation to differ from the mean, or the standard deviation

# Estimated Variance of a Random Variable

$$\text{sd}(x) = \sigma \quad = \quad \sqrt{\text{E}\left((X - \text{E}(x))^2\right)}$$

This equation defines the theoretical concept of standard deviation, but doesn't give us a formula to calculate it for a sample

# Estimated Variance of a Random Variable

$$\text{sd}(x) = \sigma \quad = \quad \sqrt{\text{E}\left((X - \text{E}(x))^2\right)}$$

This equation defines the theoretical concept of standard deviation, but doesn't give us a formula to calculate it for a sample

Here is the formula for the estimate of the standard deviation, or $\hat{\sigma}$:

$$\hat{\sigma} = \sqrt{\frac{1}{n_x - 1} \sum_{i=1}^{n_x} (x_i^2) - \frac{n_x}{n_x - 1} \bar{x}^2}$$

This calculates the amount we expect a new case to differ from the mean, adjusting for the possibility of error in our estimate of the mean itself

For a derivation, see my Stat **3**21 lecture slides

# Dispersion for a nominal variable

How do we summarize the dispersion of a nominal variable?

Range, quantiles, and variance are undefined in this case

I recommend fractionalization, which is the probability that any two observations drawn at random have the same value

Fractionalization ranges from $\epsilon > 0$ for a very dispersed variable to 1 for a nominal variable with the same value at each observation

In our wealth data, education has a fractionalization of 0.29, and race 0.59.

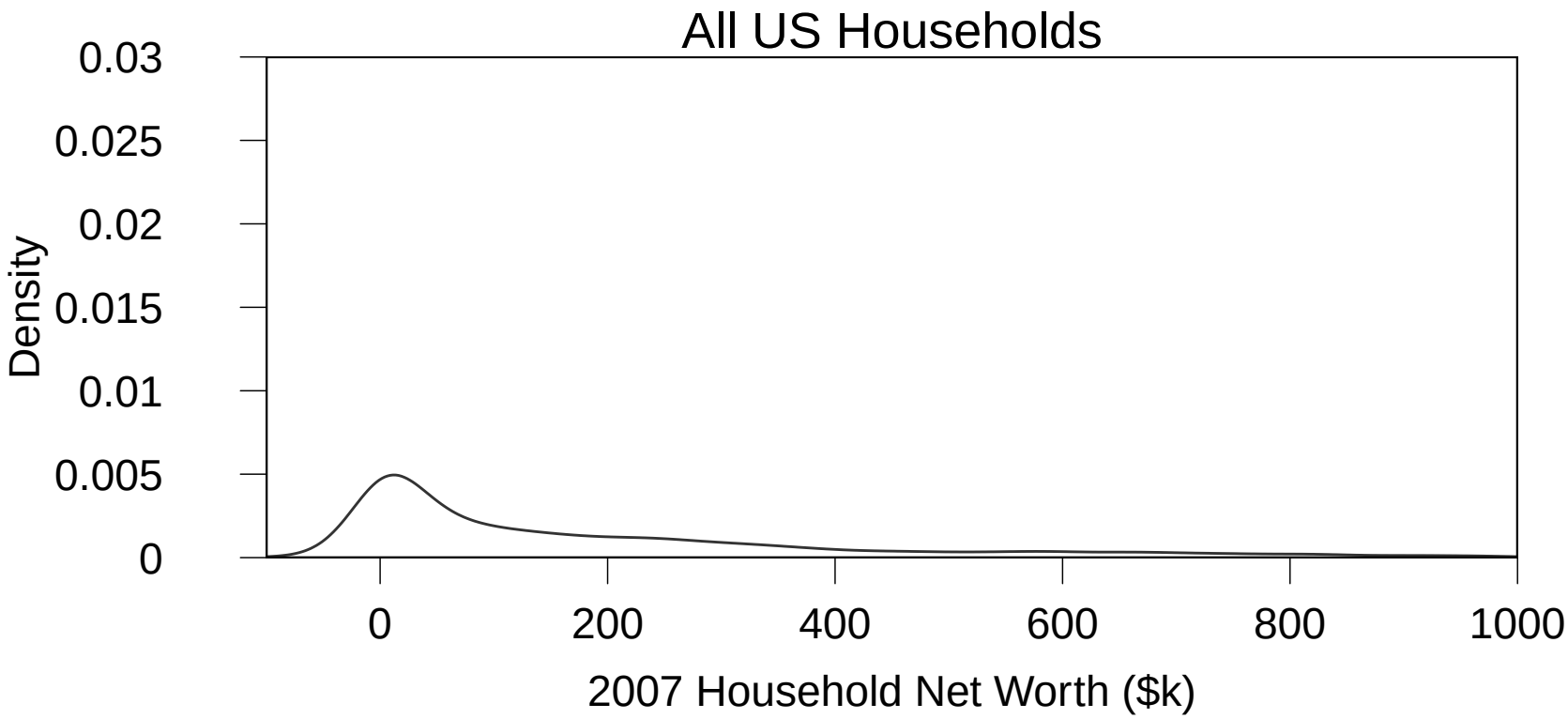
# Exploratory Data Analysis, Part 1

Visuals are as important as the “math” of data analysis

Can quickly come to grips with statistical concepts with the right pictures

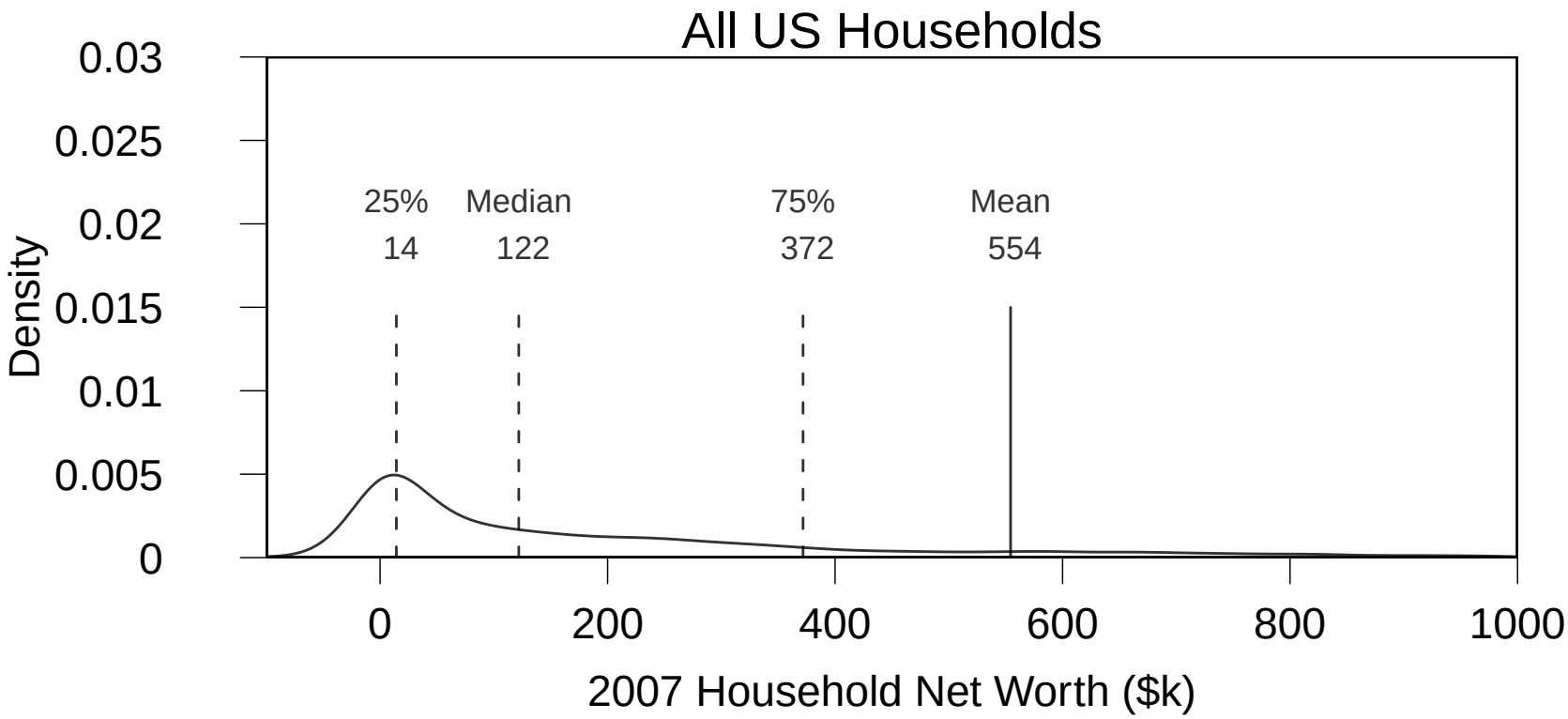
Pictures are the most powerful data analytic tool

But you need to make the right picture—and that takes statistical knowledge



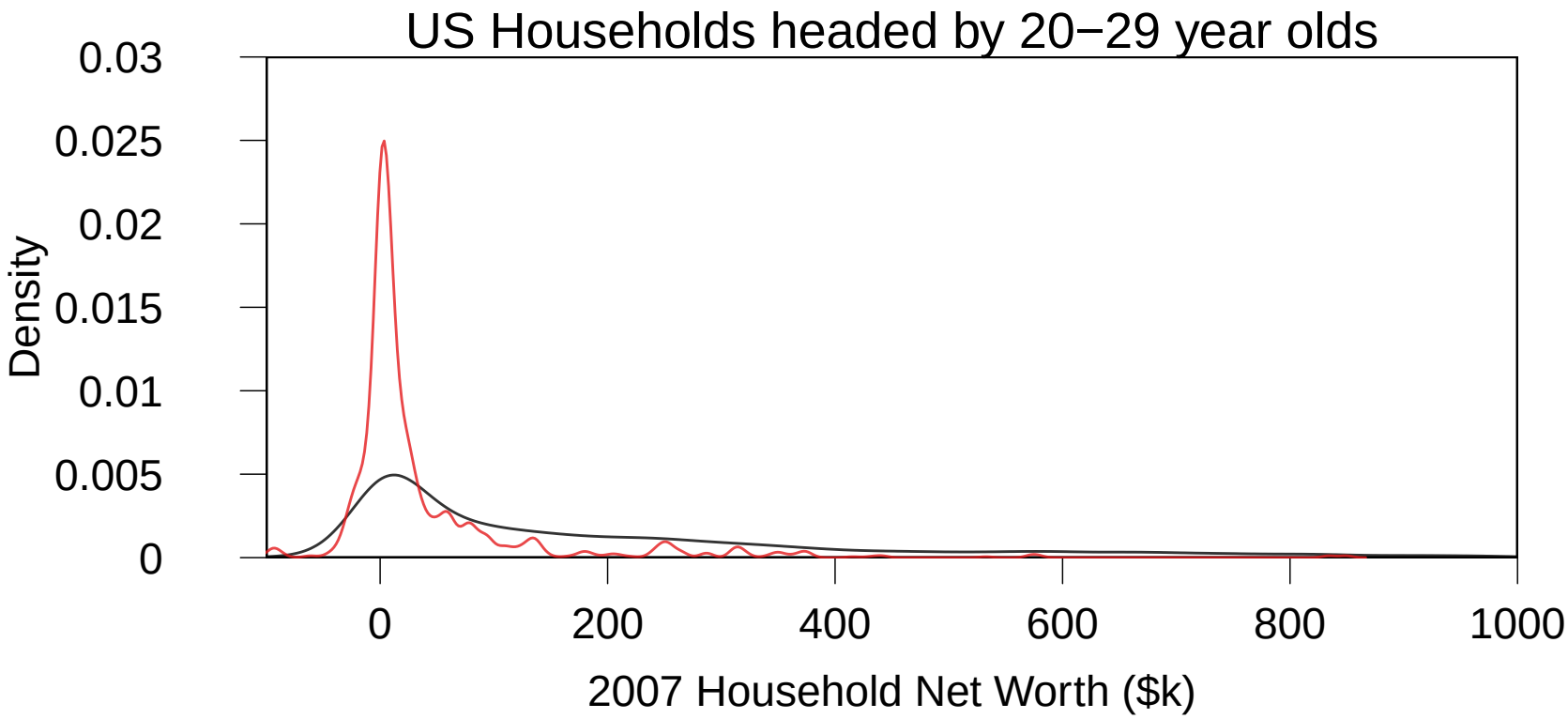

Let's explore the distribution of wealth using pictures.

We'll start by restricting the range of our density plot to the [$0, $1,000,000] interval (Why?)



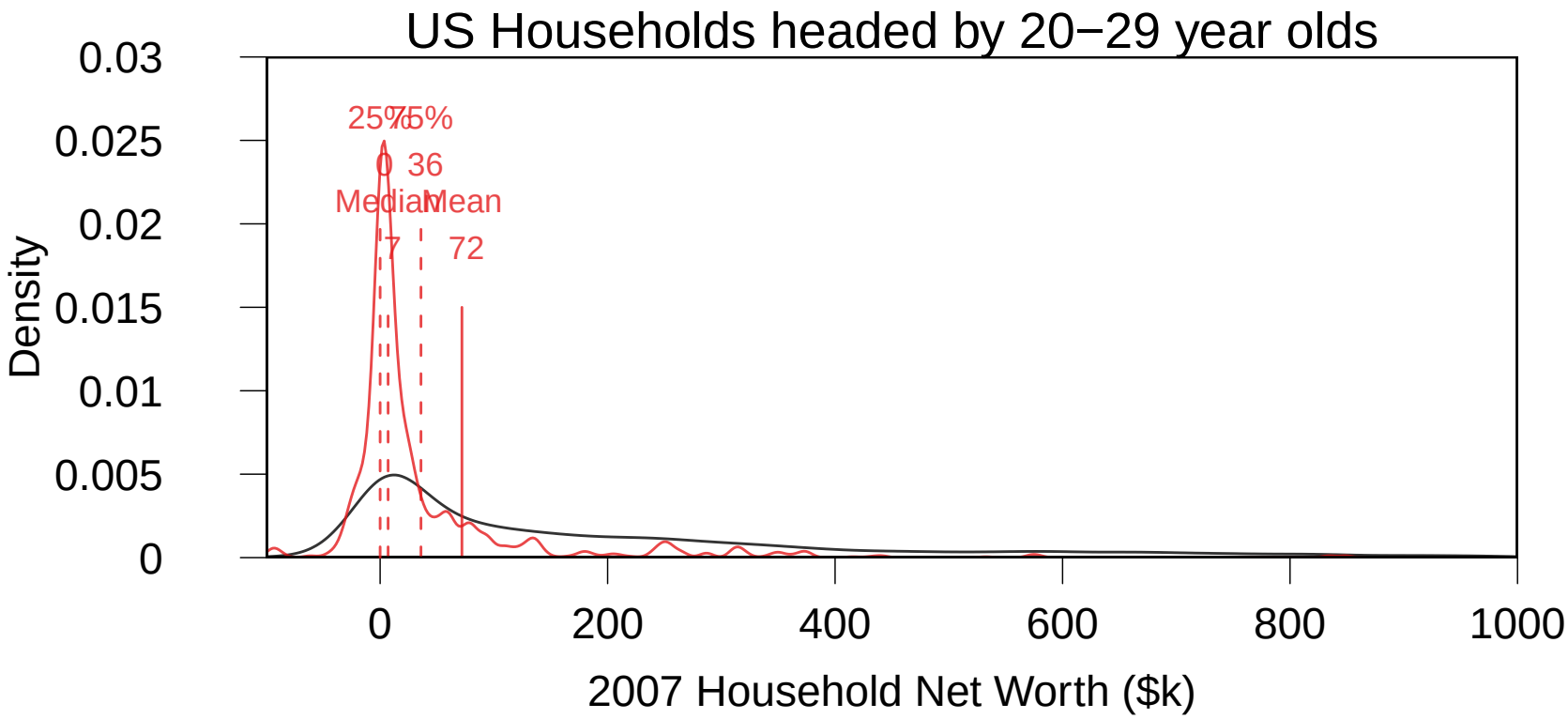

Marking the mean and interquartile range (25th and 75th percentiles) helps: note the mean is much higher than the median

This distribution is *right-skewed*: asymmetric with a long right tail



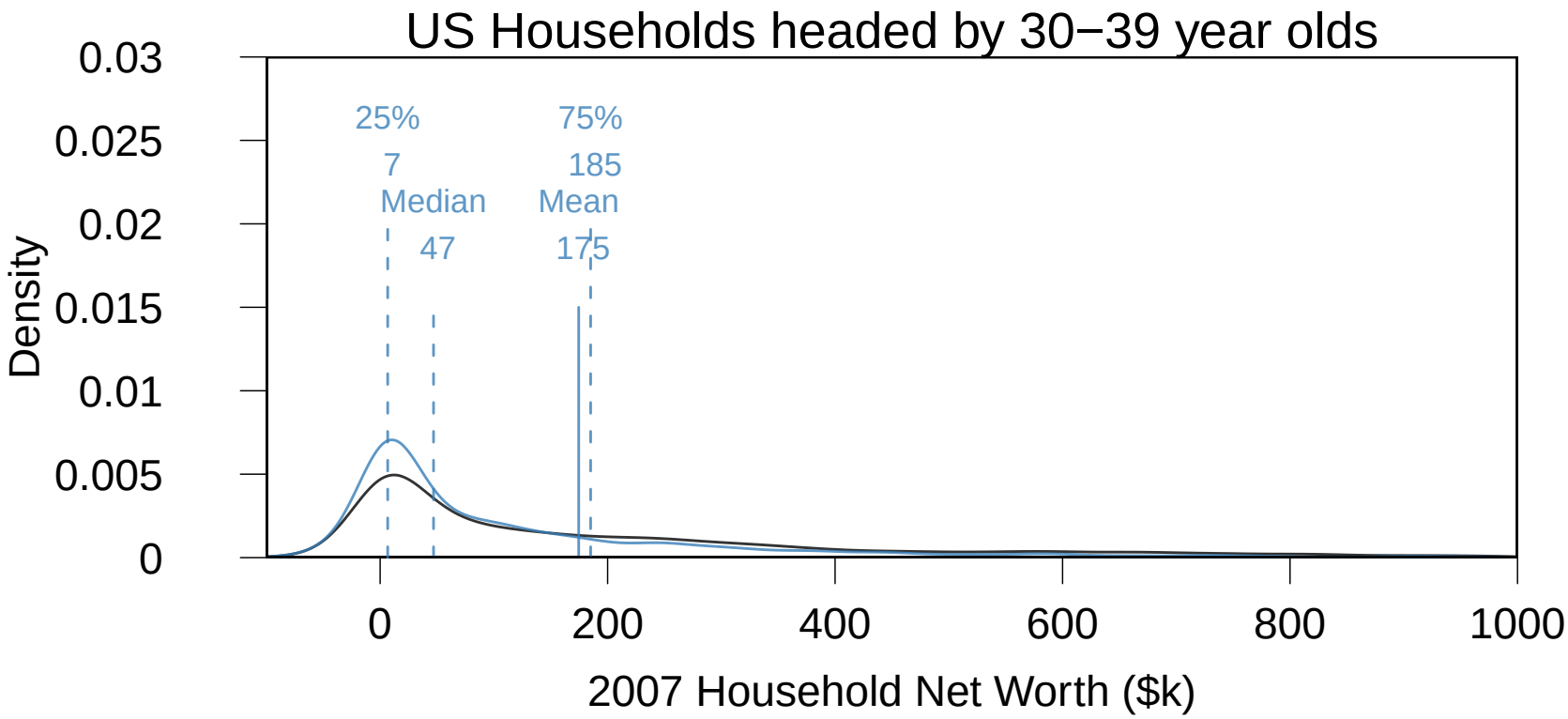

Let's select on age, to focus on people in their twenties

How do these distributions differ?



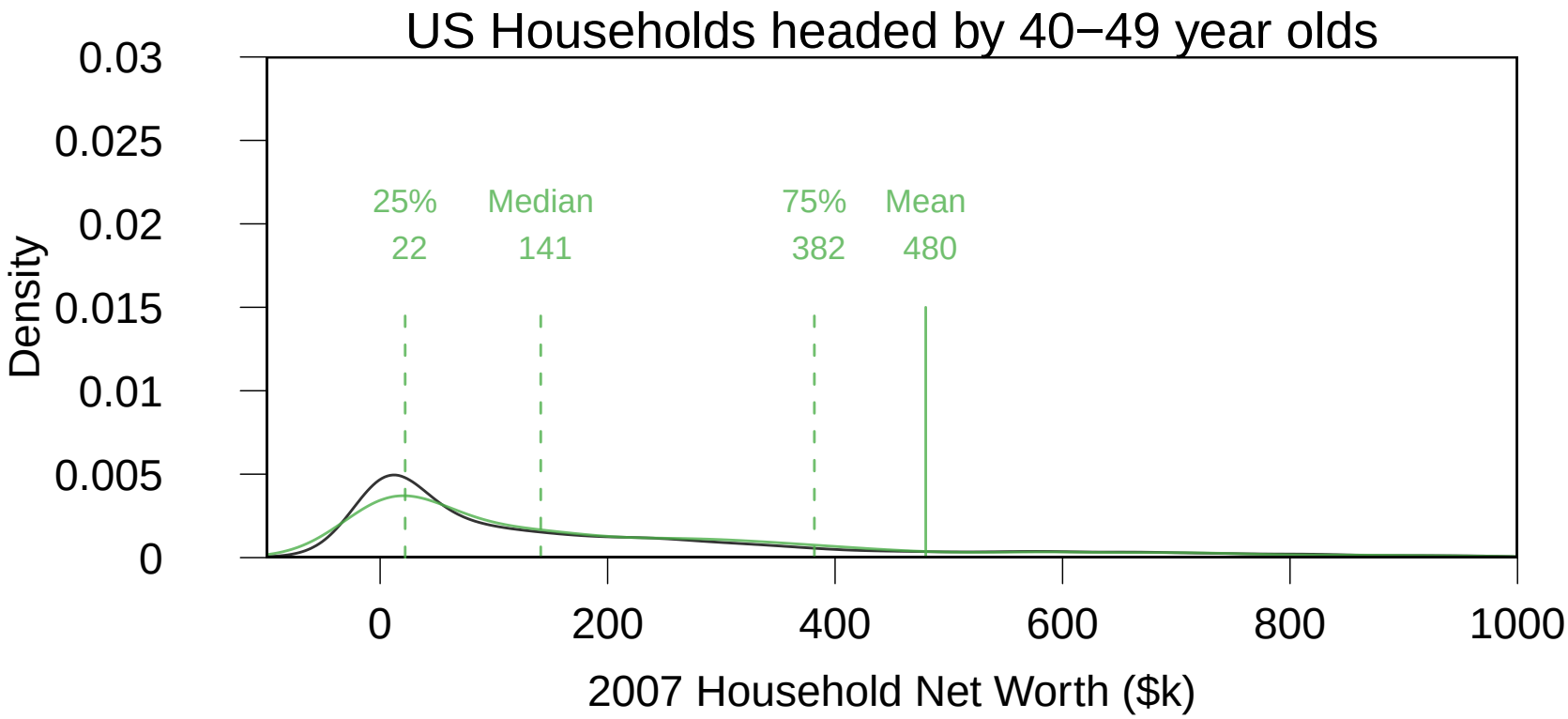

We can also compute *conditional* means and quantiles

How do these conditional measures differ from the global ones?

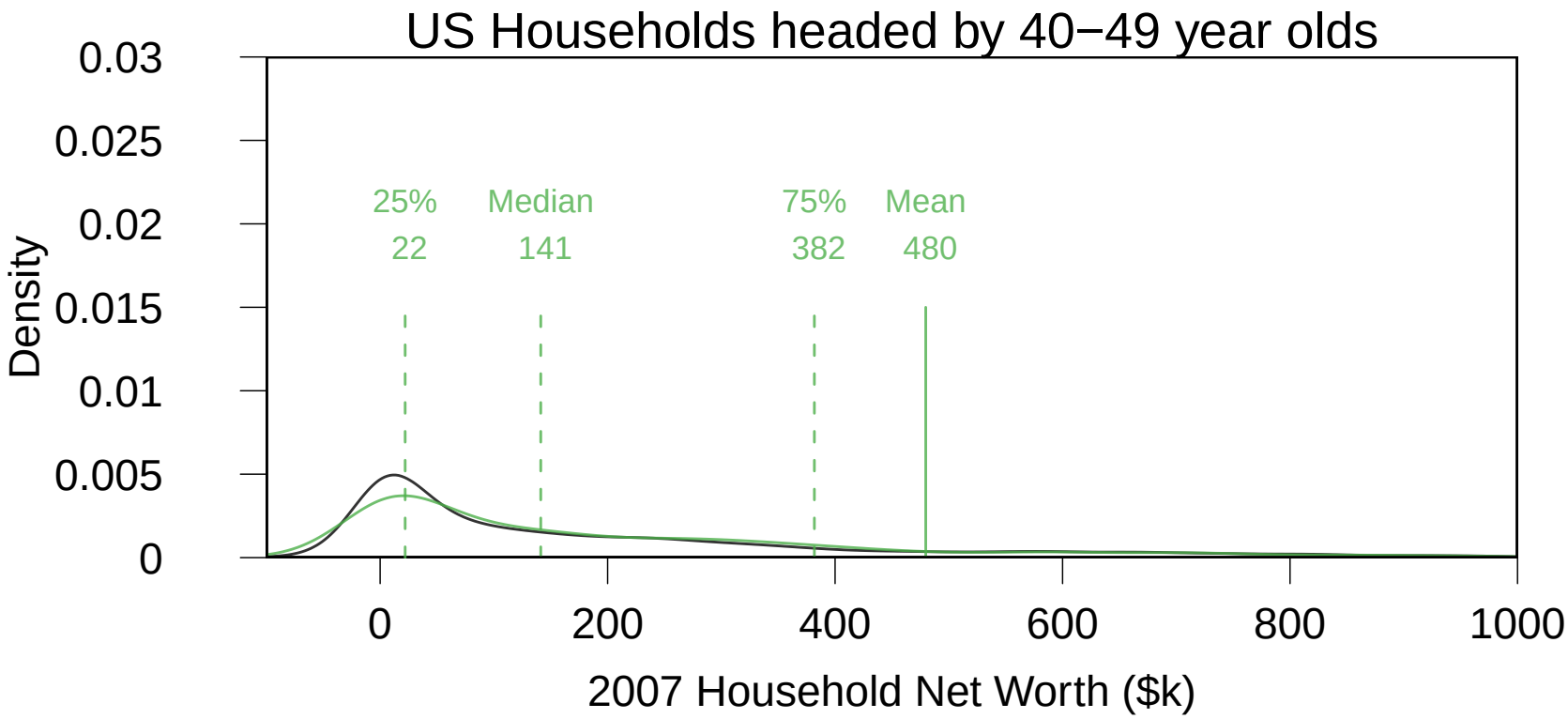



From here, we can move up the age variable by ten year increments

What do we see?

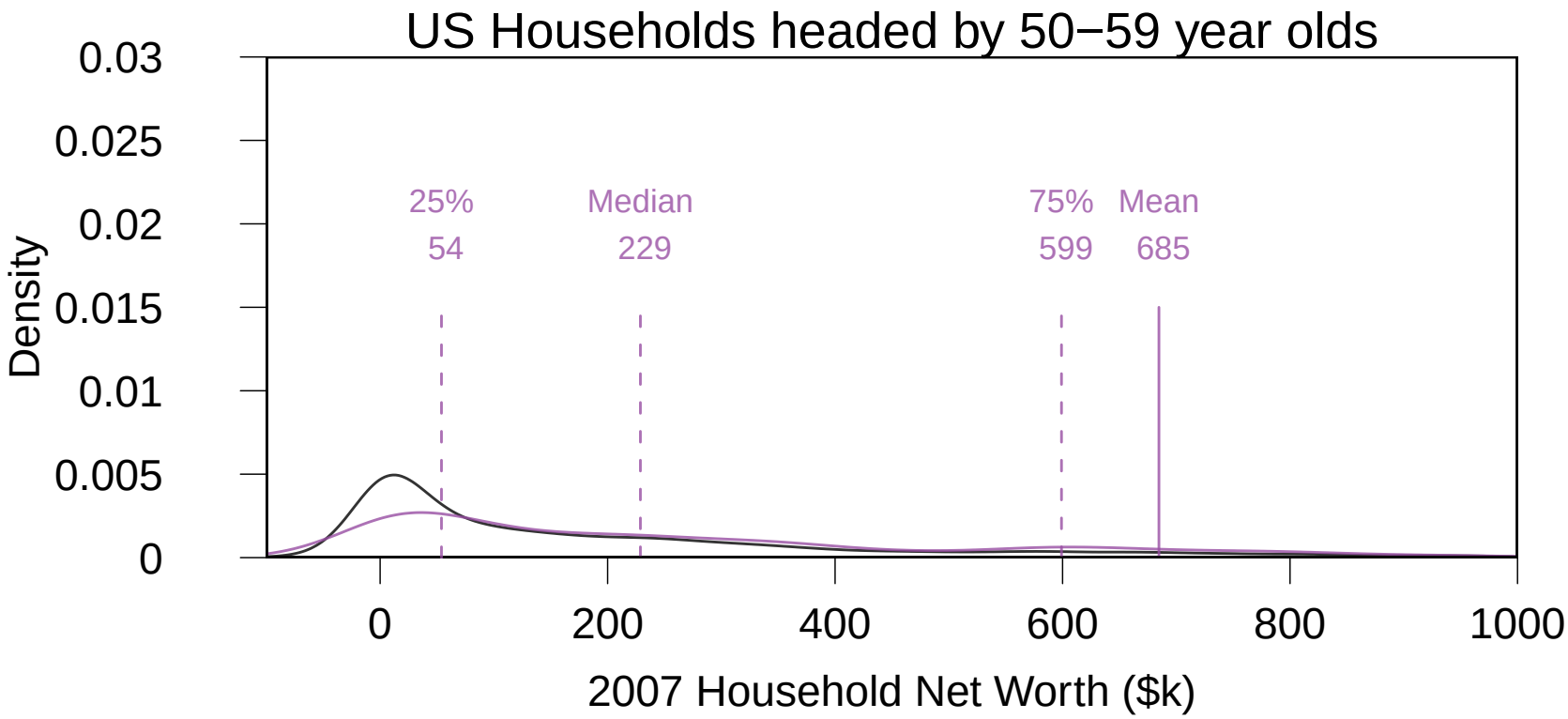



Is the apparent relationship between wealth and age necessarily causal?



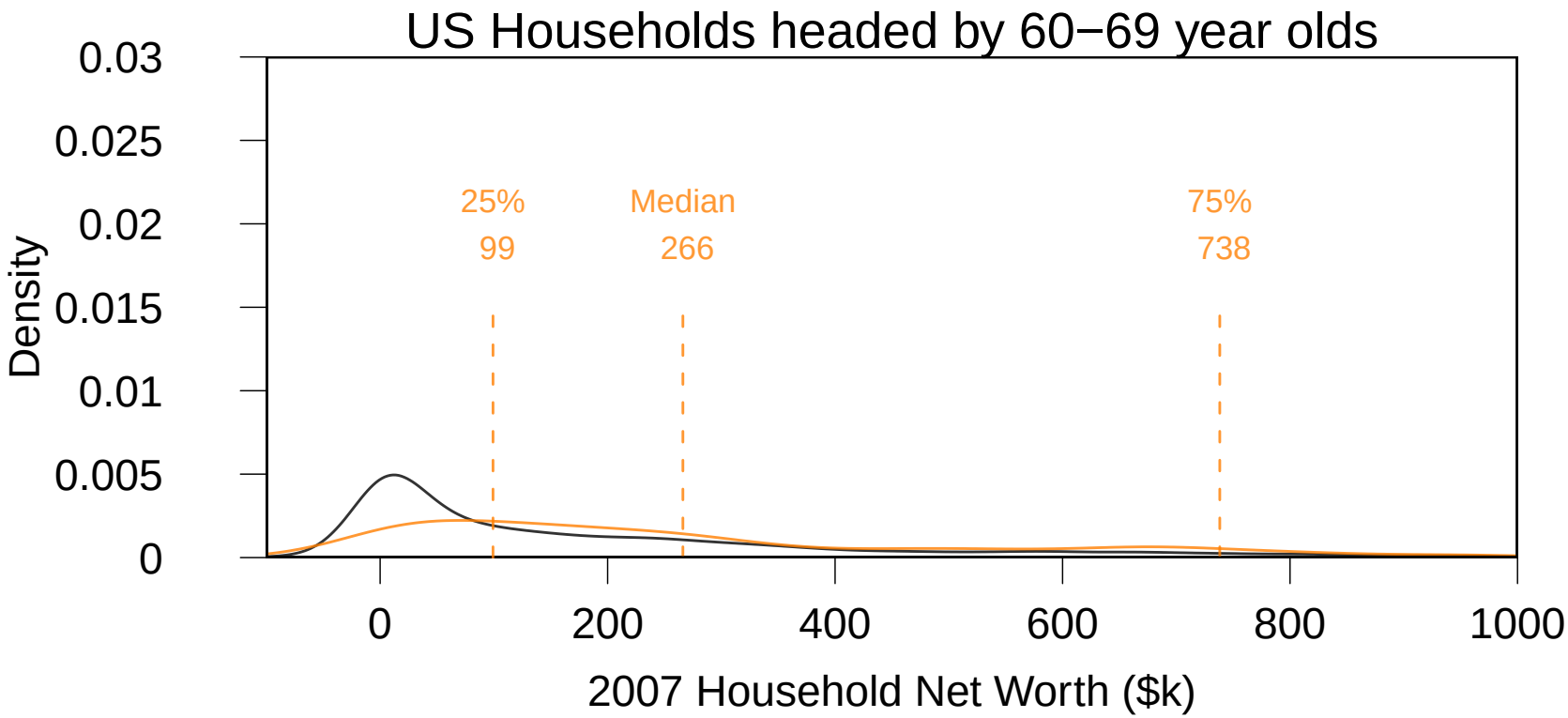

Is the apparent relationship between wealth and age necessarily causal?

What could be confounding this relationship?



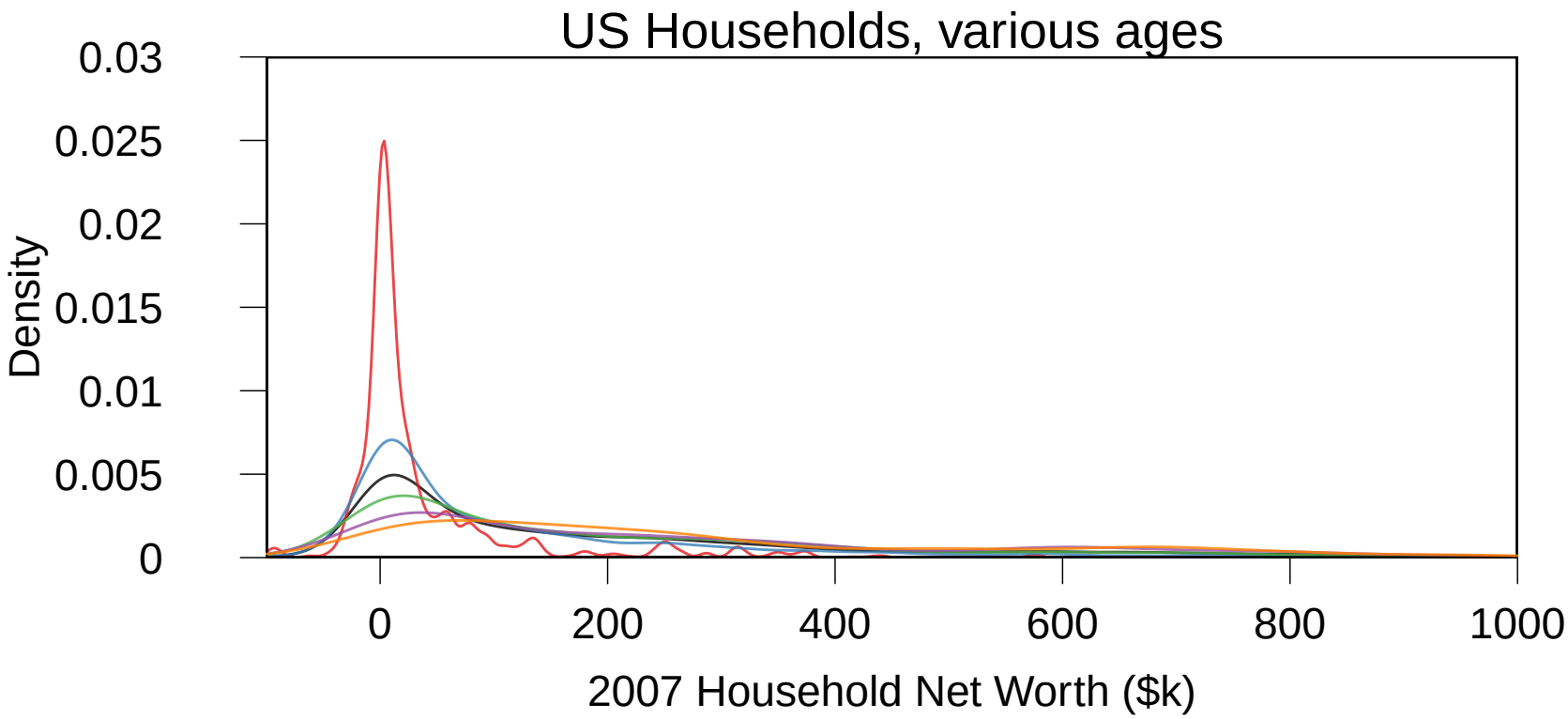

Note that more and more of our data is passing outside the right boundary with each decade

This suggests our “wealth window” approach is too limited




Indeed, the conditional mean doesn't even show up for households in their 60s!

Ideas of what to do to get the rich and not so rich on the same plot?




Looking at all the conditional plots at once

This isn't really possible with histograms—a reason to investigate density plots

# Orders of Magnitude

Notice that our wealth data range across many orders of magnitude

(Orders of magnitude means “factors of ten”)

This makes it hard to compare changes on different scales

If we scale our graph from \$0 to \$100 million,
the range from 0 to \$1 million almost disappears in the left corner

But that’s where most people are!

# Logarithms

Logarithms help us look at the small and large scale in the same graph

A logarithm is just a mathematical function, $\log(x)$, which replaces variables with their order of magnitude

Specifically, the $\log_a(x)$, or "log base $a$ of $x$", is defined:

$$a^{\log_a(x)} = x$$

That is, when $a$ is raised to this number, it equals $x$

# The natural logarithm

When taking logs, does it matter which base we use?

Not really

The base $a$ can be any positive number,
but is most commonly 10 or a special number $e$

$e$ is a famous constant, $2.718281828\ldots$

Logs in base $e$ are known as natural logarithms

In this course, $\log(x)$ will be understood to mean $\log_e(x)$

# Logarithms

| $x$ | $\log(x)$ | $\log_{10}(x)$ |
|---|---|---|
| 0.0001 | −9.21 | −4 |
| 0.001 | −6.91 | −3 |
| 0.01 | −4.61 | −2 |
| 0.1 | −2.30 | −1 |
| 1 | 0.00 | 0 |
| 10 | 2.30 | 1 |
| 100 | 4.61 | 2 |
| 1000 | 6.91 | 3 |
| 10000 | 9.21 | 4 |
| 100000 | 11.51 | 5 |

Note that in each row, the $\log_{10}$ and the natural log differed only by a constant factor, 2.30.

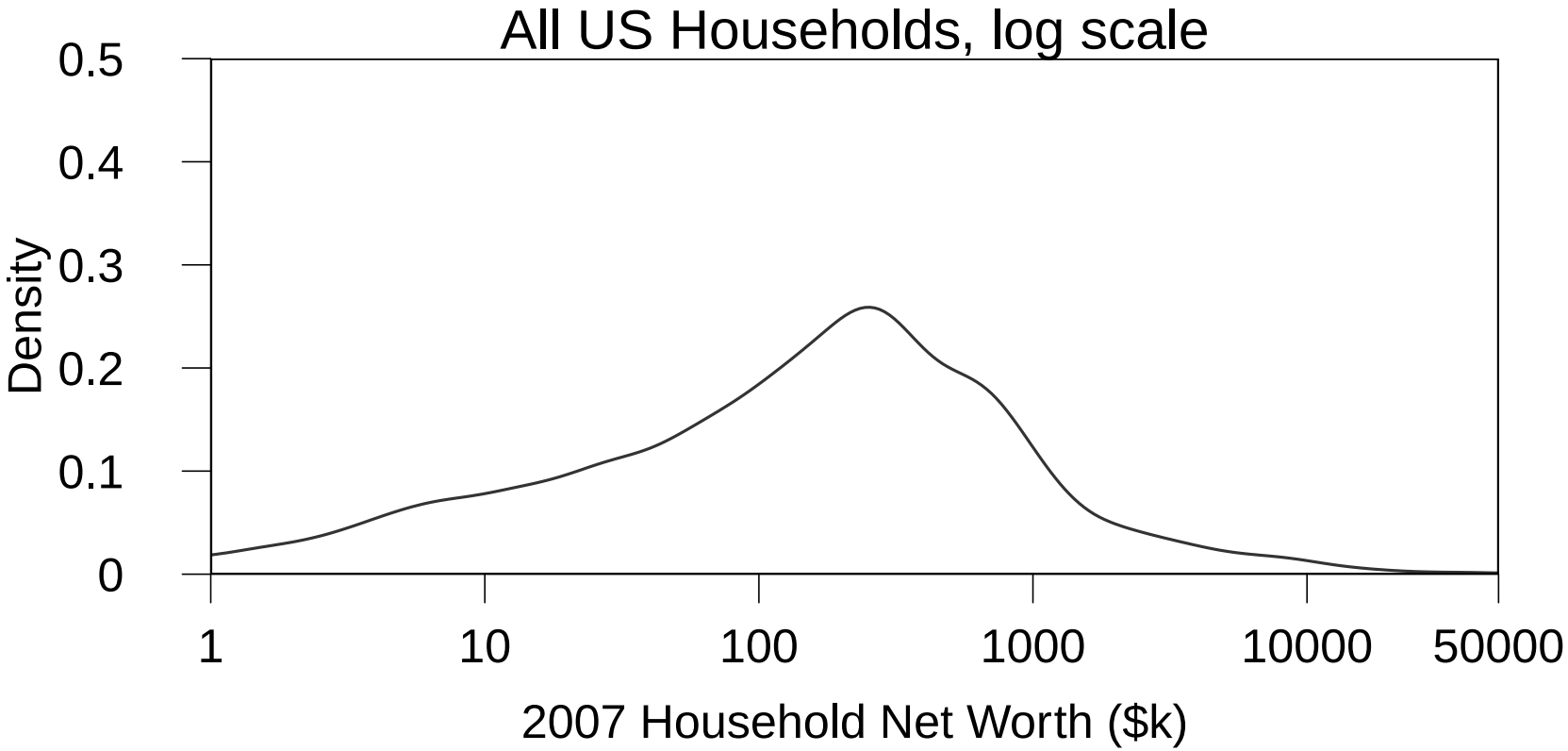
All bases of other logarithms differ from the natural log only by a constant multiplicative factor

Finally, note that we can only take the log of positive numbers.

# Exponentiation

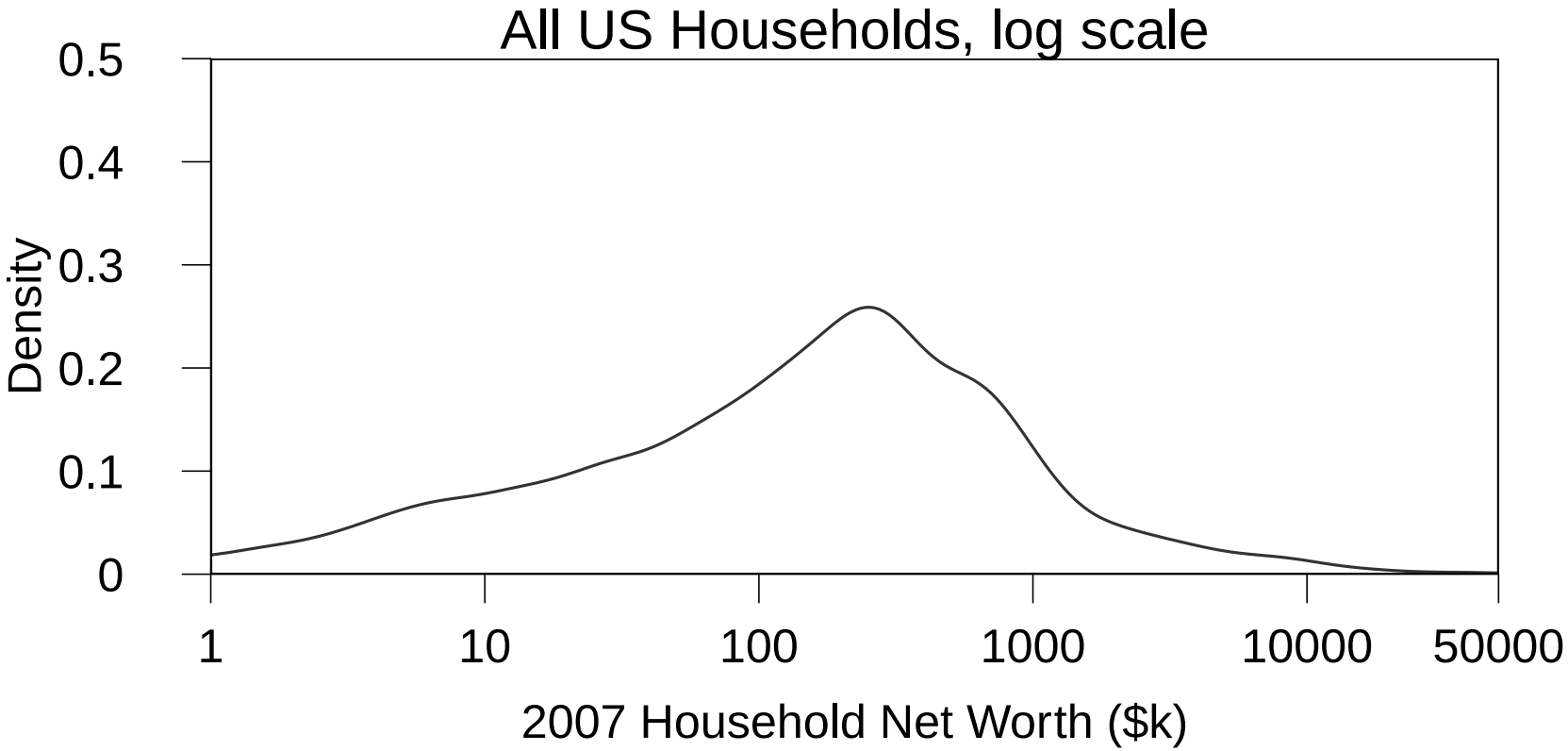
The inverse of the logarithm is just $a^x$, or exponentiation.

We will write $e^x$ as $\exp(x)$ to avoid confusion



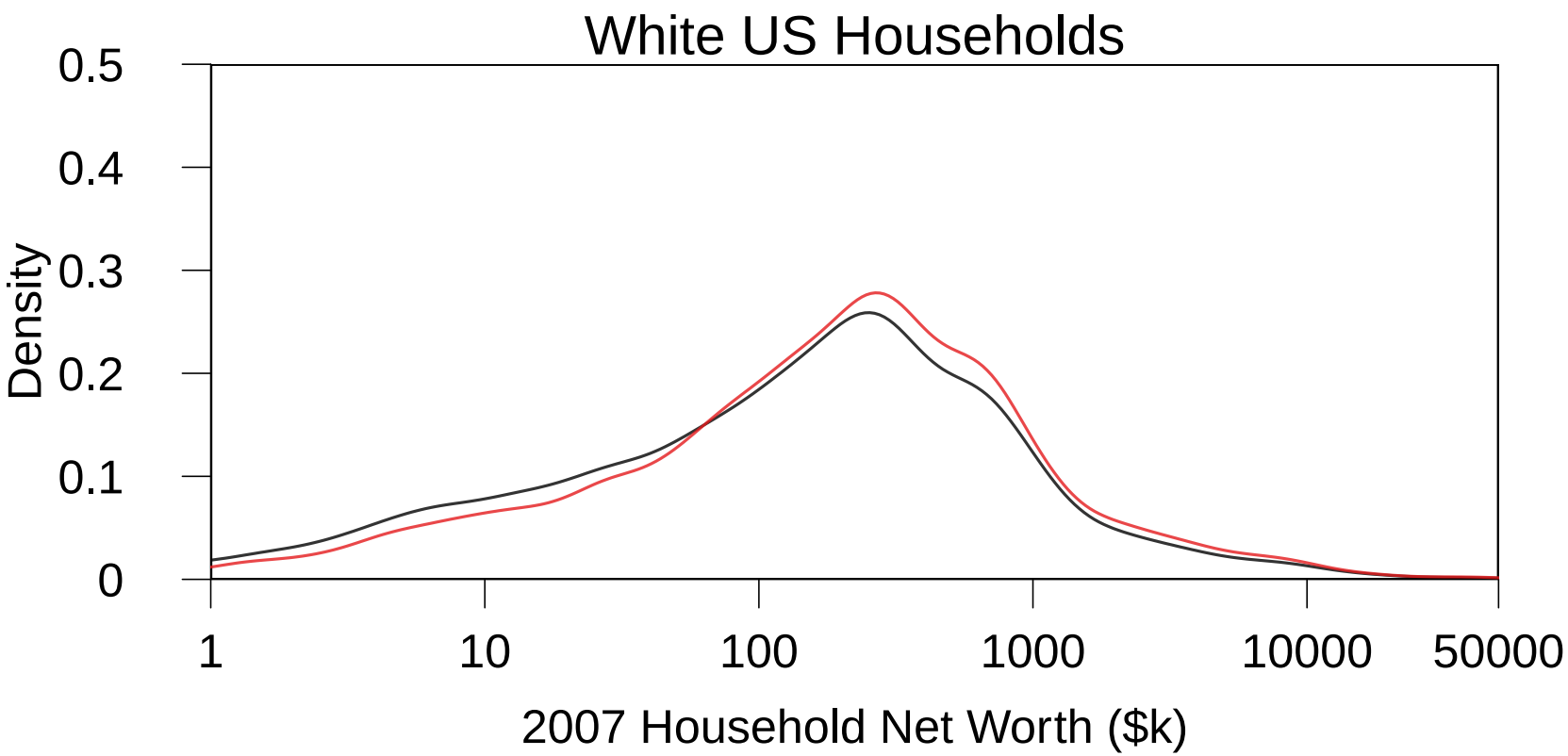

Logging wealth before plotting yields allows us to plot the whole positive range of wealth

We have to leave out negative wealth because negative values can't be logged



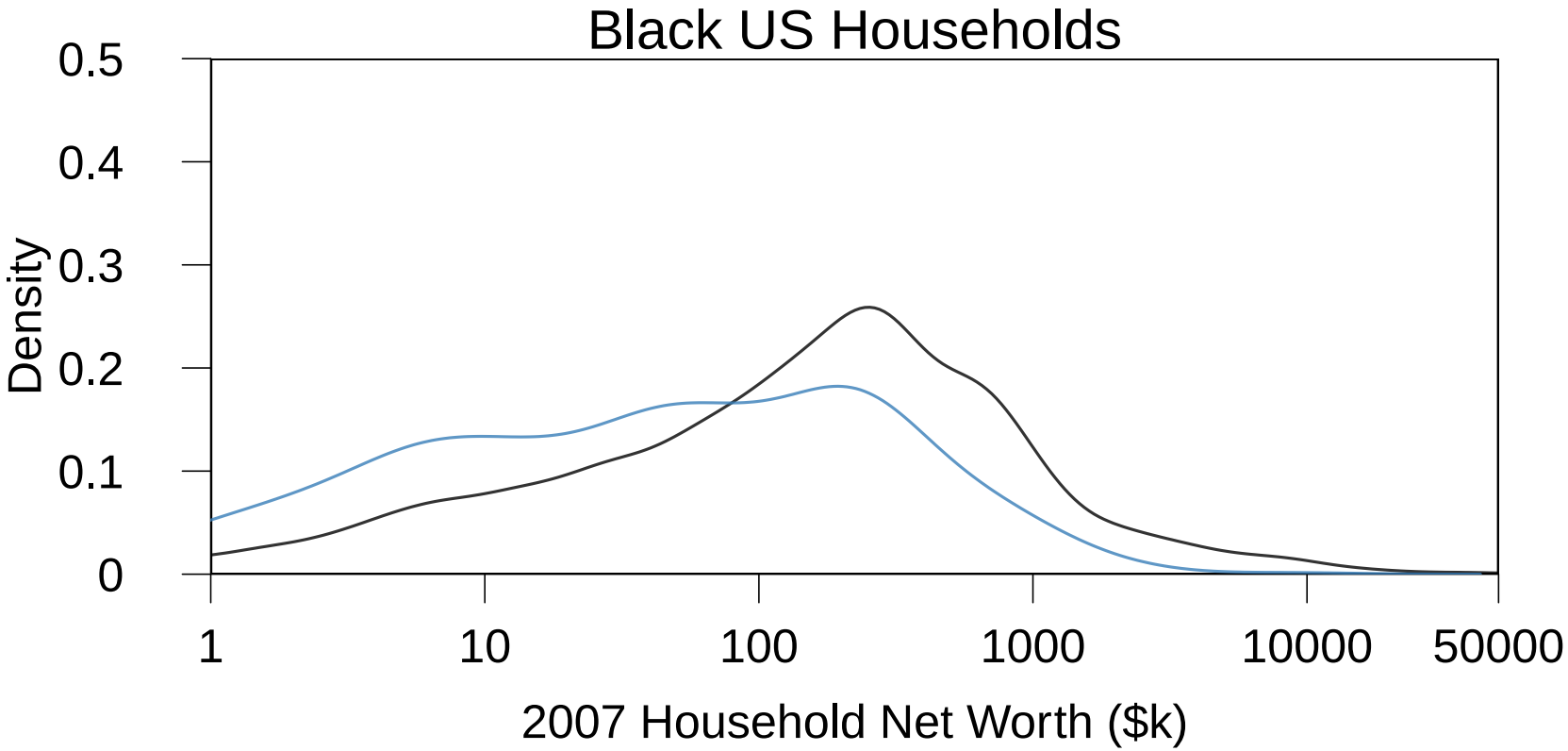

This way of plotting is also referred to as “logging the axis”

That’s because all we’ve done is squeezed the horizontal axis more and more as we move to the right



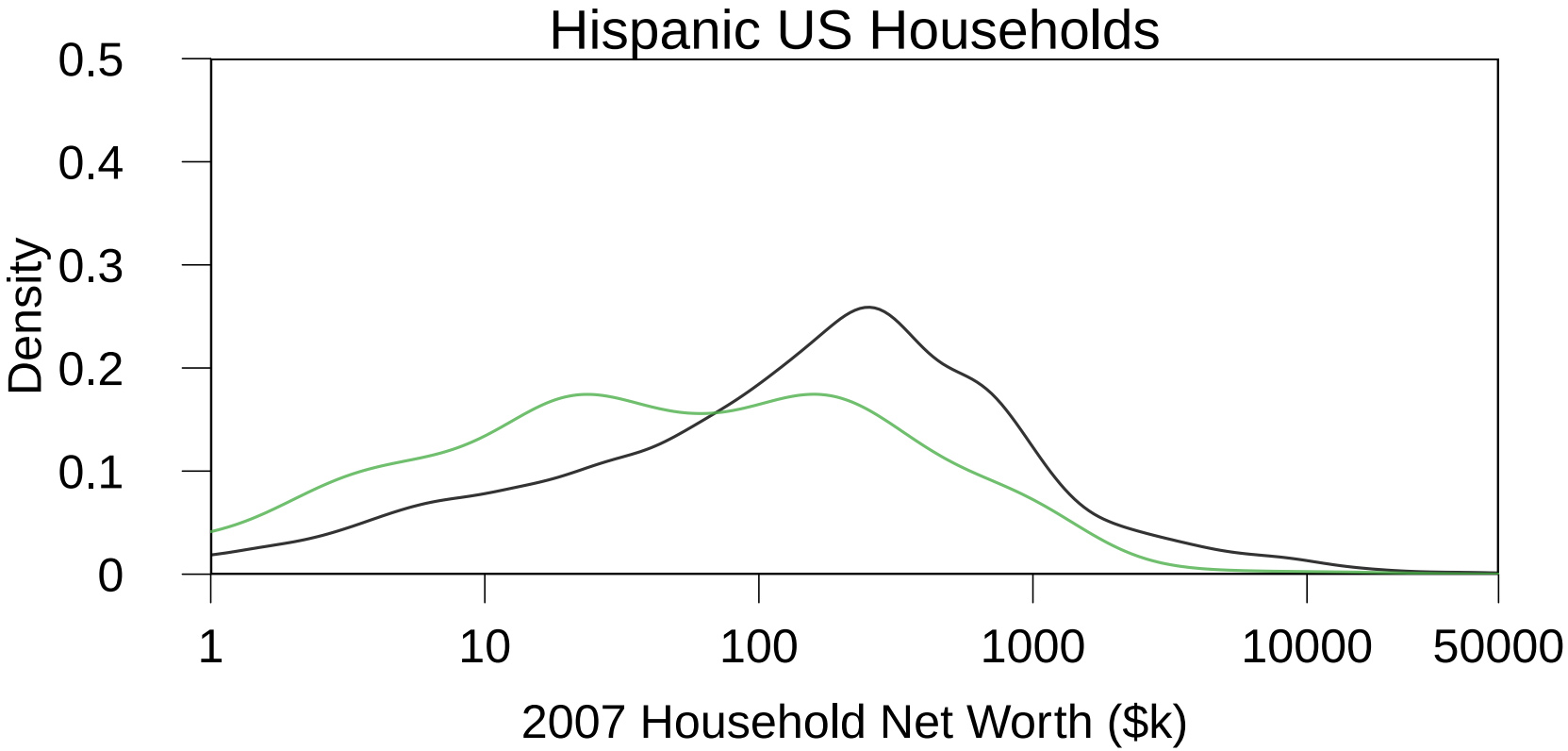

Let's use this new tool to compare wealth for different self-identified racial groups

Here are households with self-identified white heads



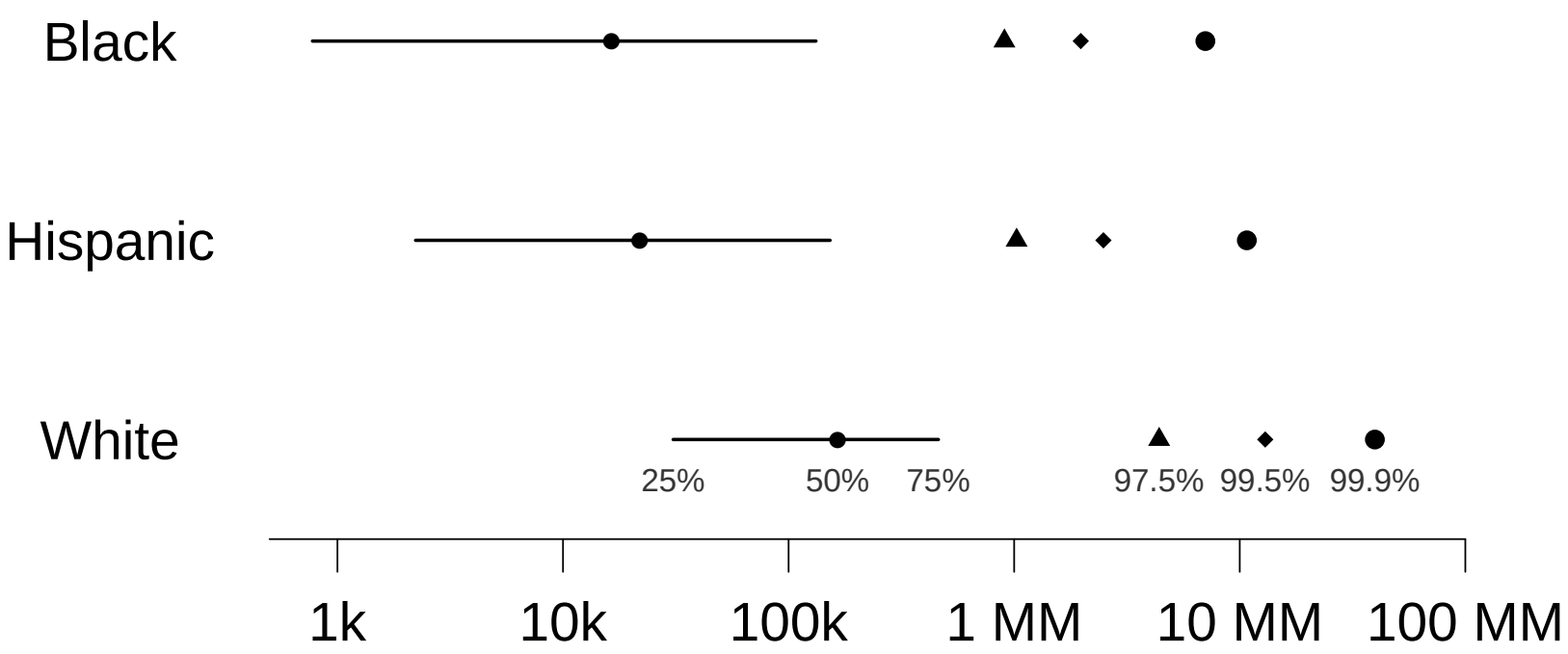

Logging helps reveal the dramatic racial differences at both ends of wealth spectrum

Can we attribute these differences to a direct causal relationship with race (ie, racial discrimination?)



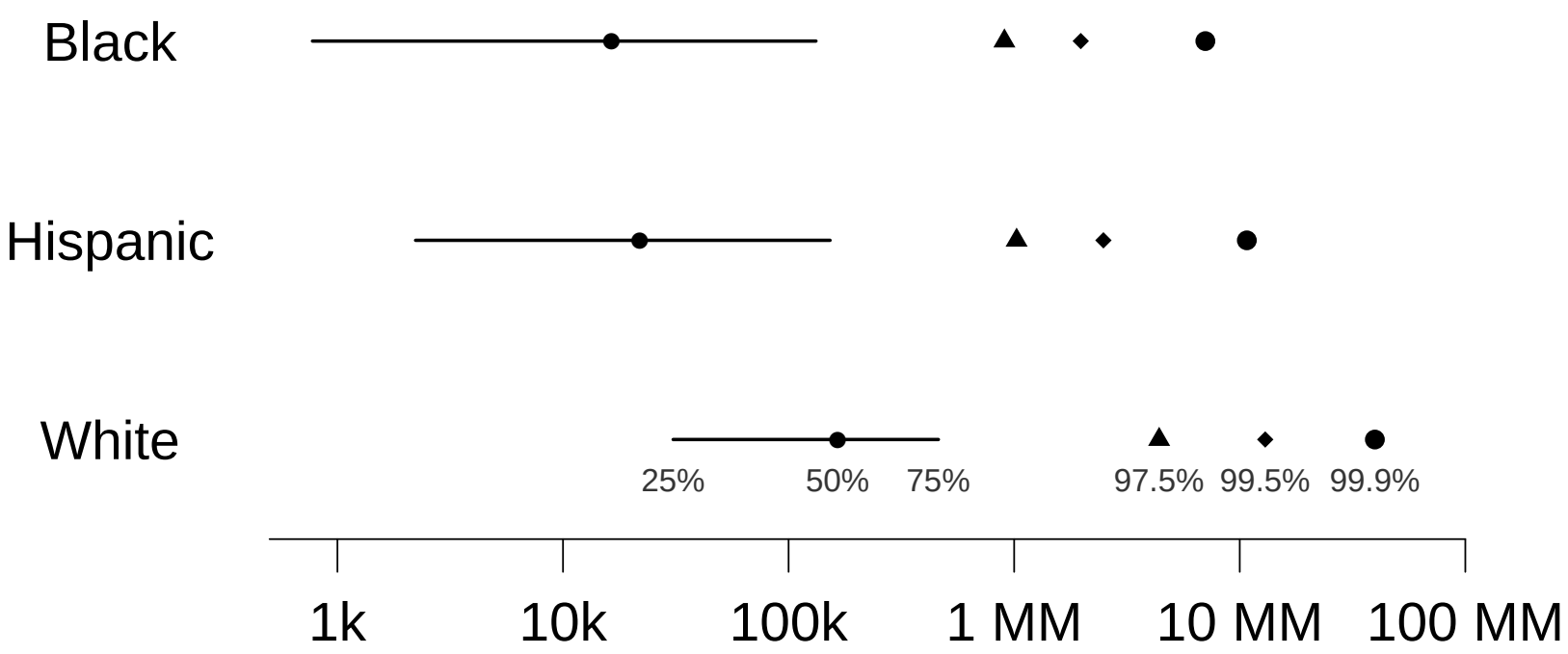

What can we say about the three-way comparison of whites, blacks, and Hispanics?

It would be nice to summarize the key points in a single graphic



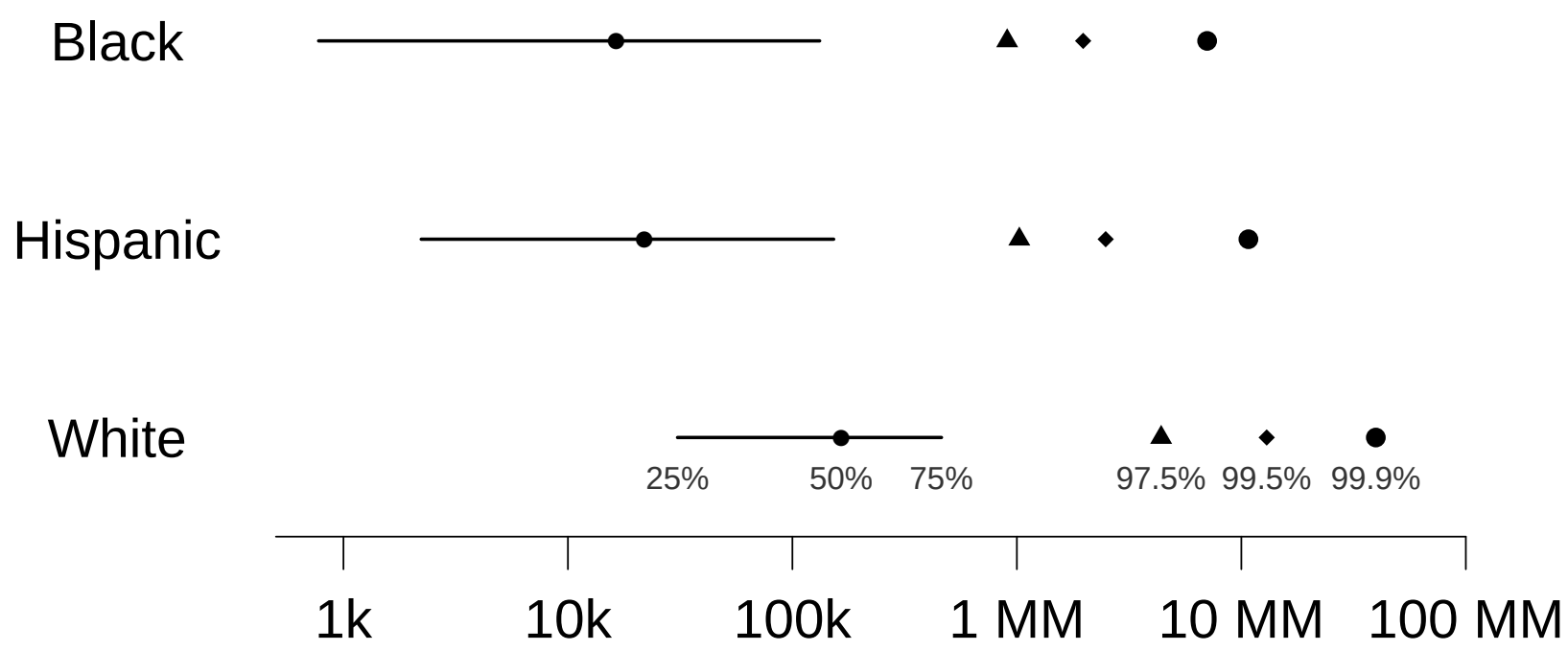

The above plot is known as a dotplot.

Similar to a boxplot (more on boxplots later)



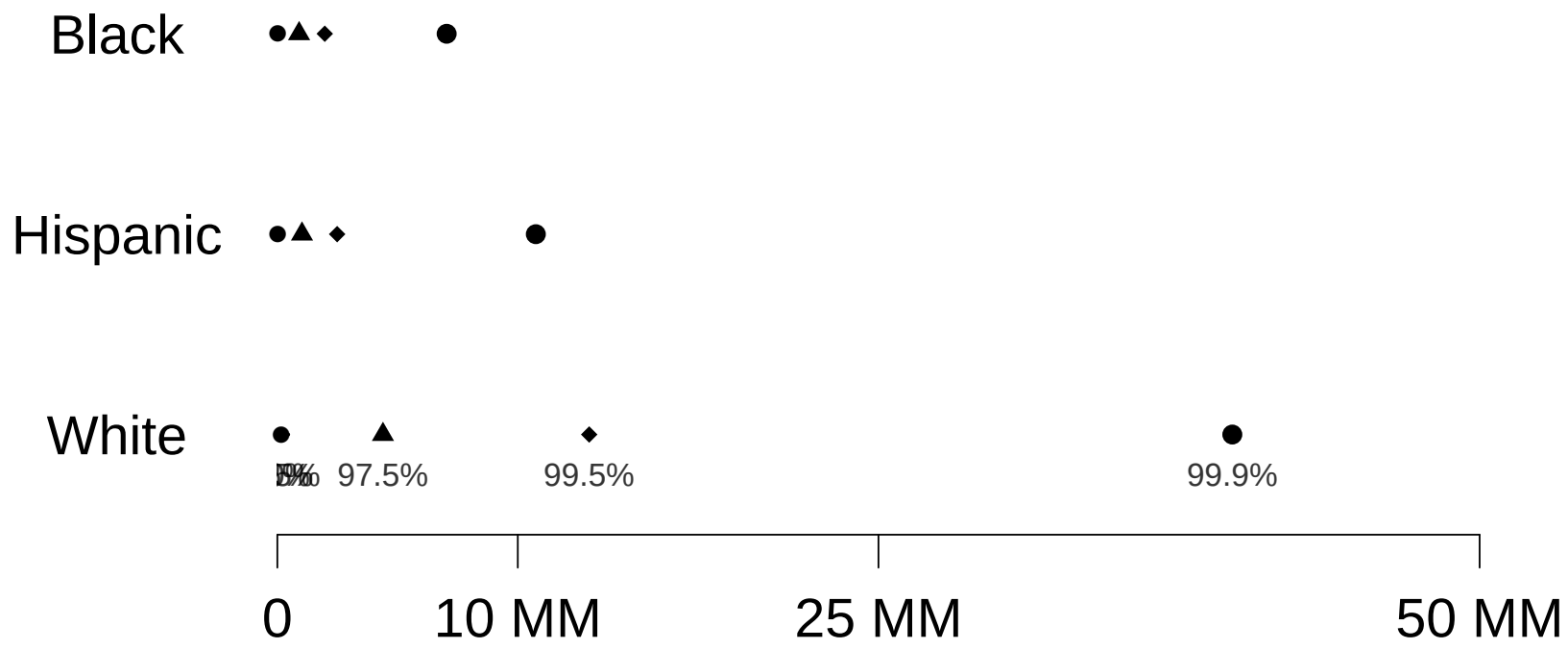

Not as much detail as histograms or densities, but a good compromise for side-by-side comparison.

All good data graphics facilitate comparison of variables or their conditional central tendencies or variances



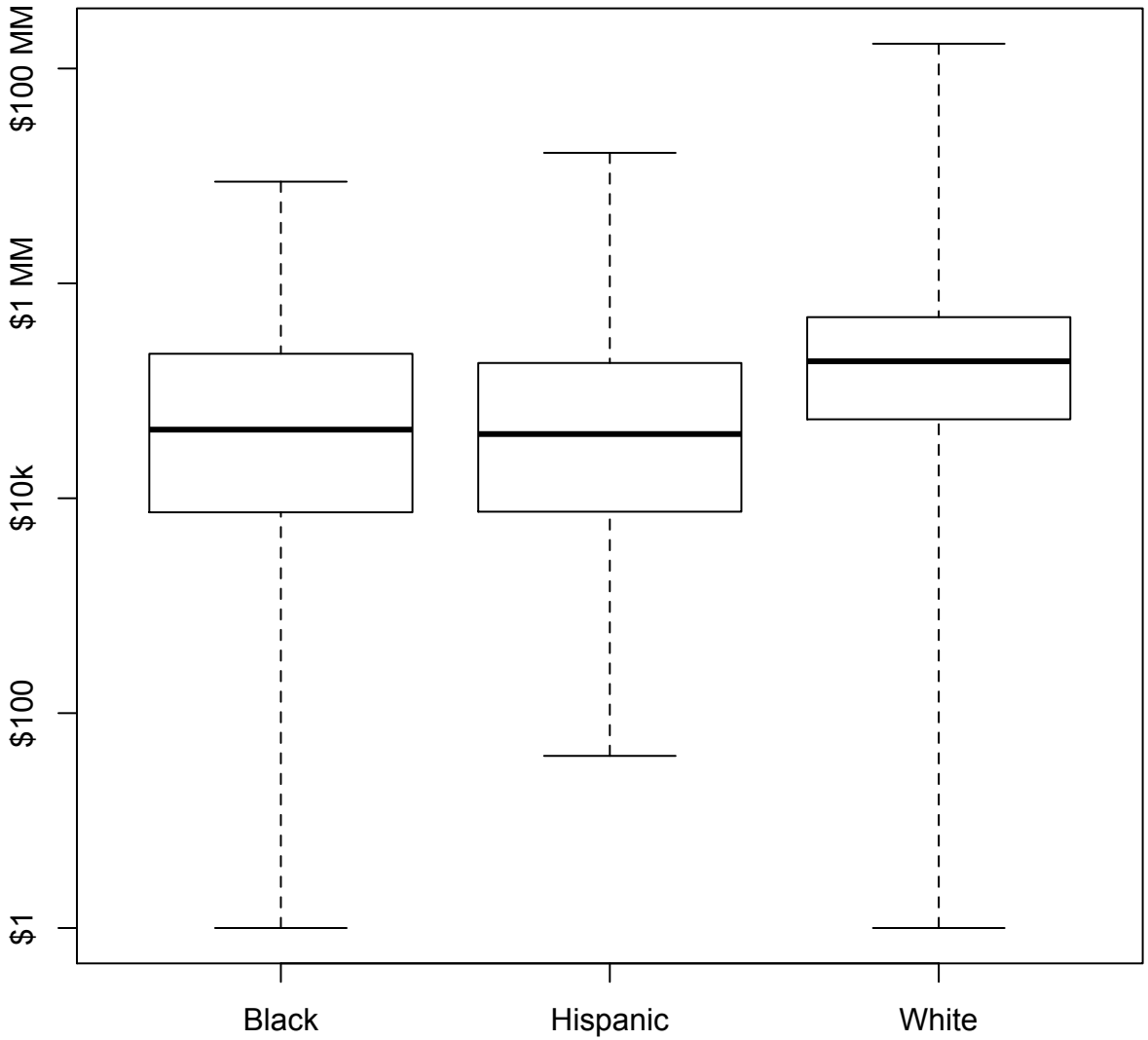

Now clear why the “Occupy Wall Street” movement has connected with many

Even before the recession (2007 data),
a huge gap between the middle class (middle 50%) and the top 1%




Why logging is important:
Unlogged, almost every household in the US vanishes into a single dot

You won't always log your variables,
but if you can "count" it, you probably should try logging it




A common and useful alternative: the boxplot

Shows the "5-number summary":

max
75th percentile
median
25th percentile
min