Lecture 2: Statistics for 1 Variable

Intro Stats with Nathan Favero

American University (Washington, DC)

August 2, 2024

Measures of central tendency

- Q: What is the typical value for this variable? (Can only answer with one number; can't give me a range)
- A: We want a measure of central tendency

Measures of central tendency

- Mean: average; add up all values and then divide by total number of observations

$$\mu = \frac{\sum x_i}{N}$$

- Median: 50th percentile; line all values up from least to greatest and then pick the value in the very middle. If two values are tied, take the average of the two.
- Mode: most frequently occurring value

Measures of central tendency

- Mean: sensitive to outliers (unusually large or small values)
- Median: not sensitive to outliers
- Mode
  - Usually not ideal for continuous variables (especially when measured to many digits), although can use bins instead with continuous variables
  - There may be more than one mode

Measures of spread

- Q: How tightly is the data clustered around the center of the distribution? Are most data points pretty close to the mean/median?
- A: We want a measure of spread

Measures of spread

- Range: maximum value minus minimum value
  - Doesn't necessarily tell us anything about whether most observations are close to the mean/median
- Interquartile range (IQR): 75th percentile minus 25th percentile
  - 25th and 75th percentile calculated like median, except they mark the 25% point and 75% point (as opposed to the middle, or 50% point) when data is arranged from least to greatest
  - IQR is depicted in a bloxplot

Measures of spread

- Most common measures of spread: variance and standard deviation
- Variance

$$\sigma^2 = \frac{\sum(x_i - \mu)^2}{N}$$

  - This is equal to the average squared deviation (or squared error)—in other words, the average squared difference between each data point and the mean

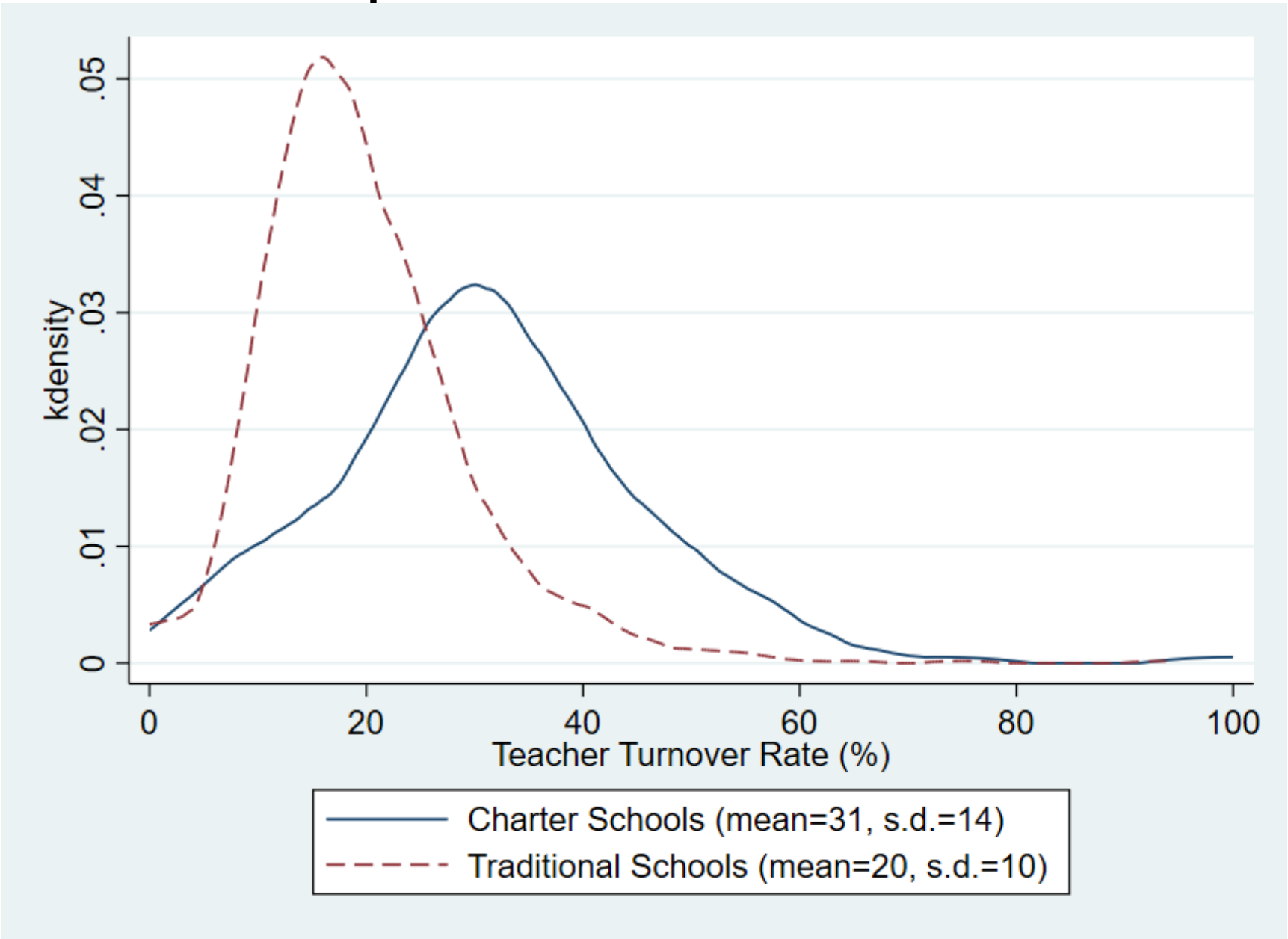
Measures of spread

- Standard deviation

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$$

  - This is just the square root of the variance
  - You can think of it as the typical distance between a data point and the mean (not quite right but a good approximation)

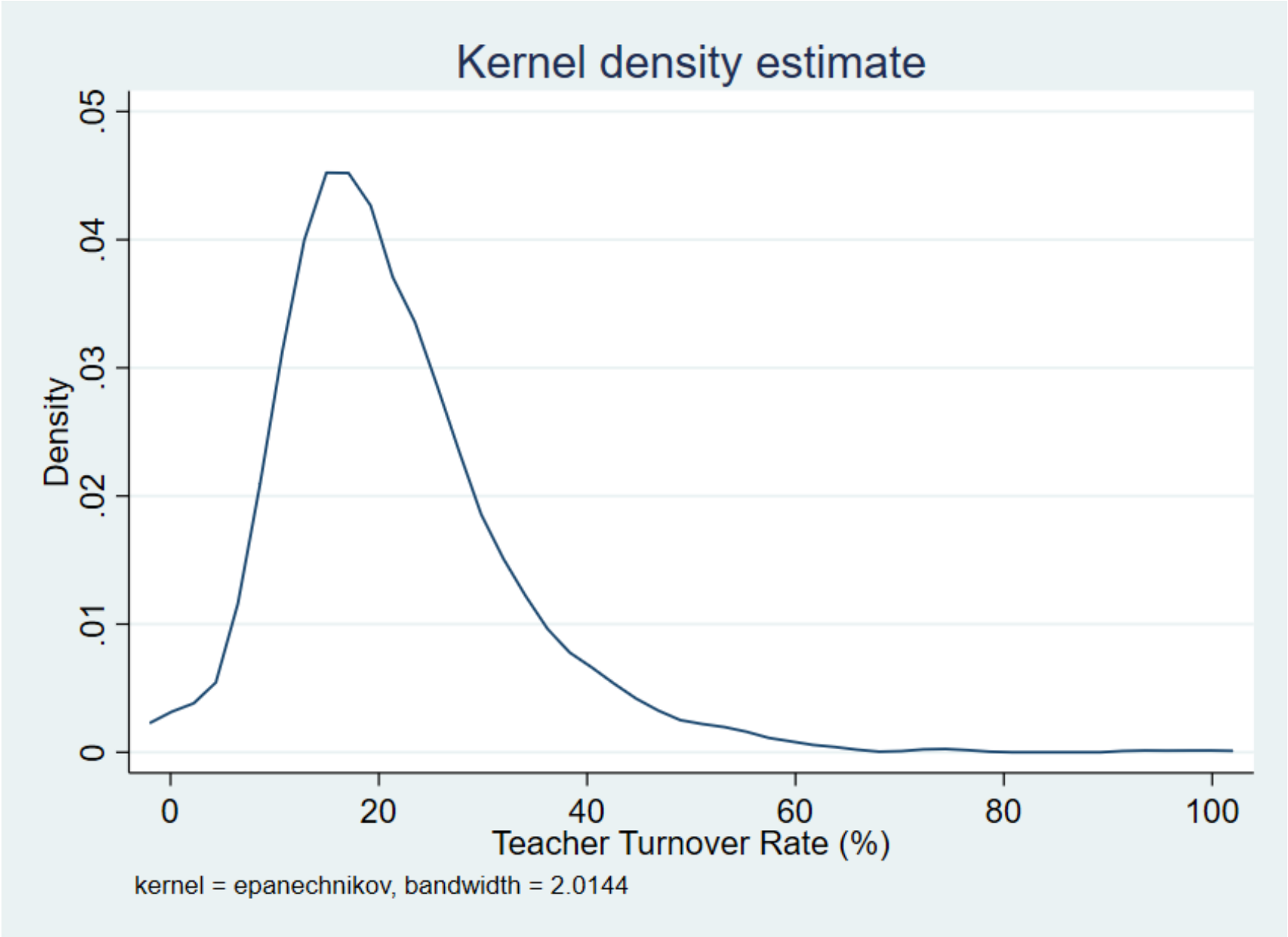
Measures of spread



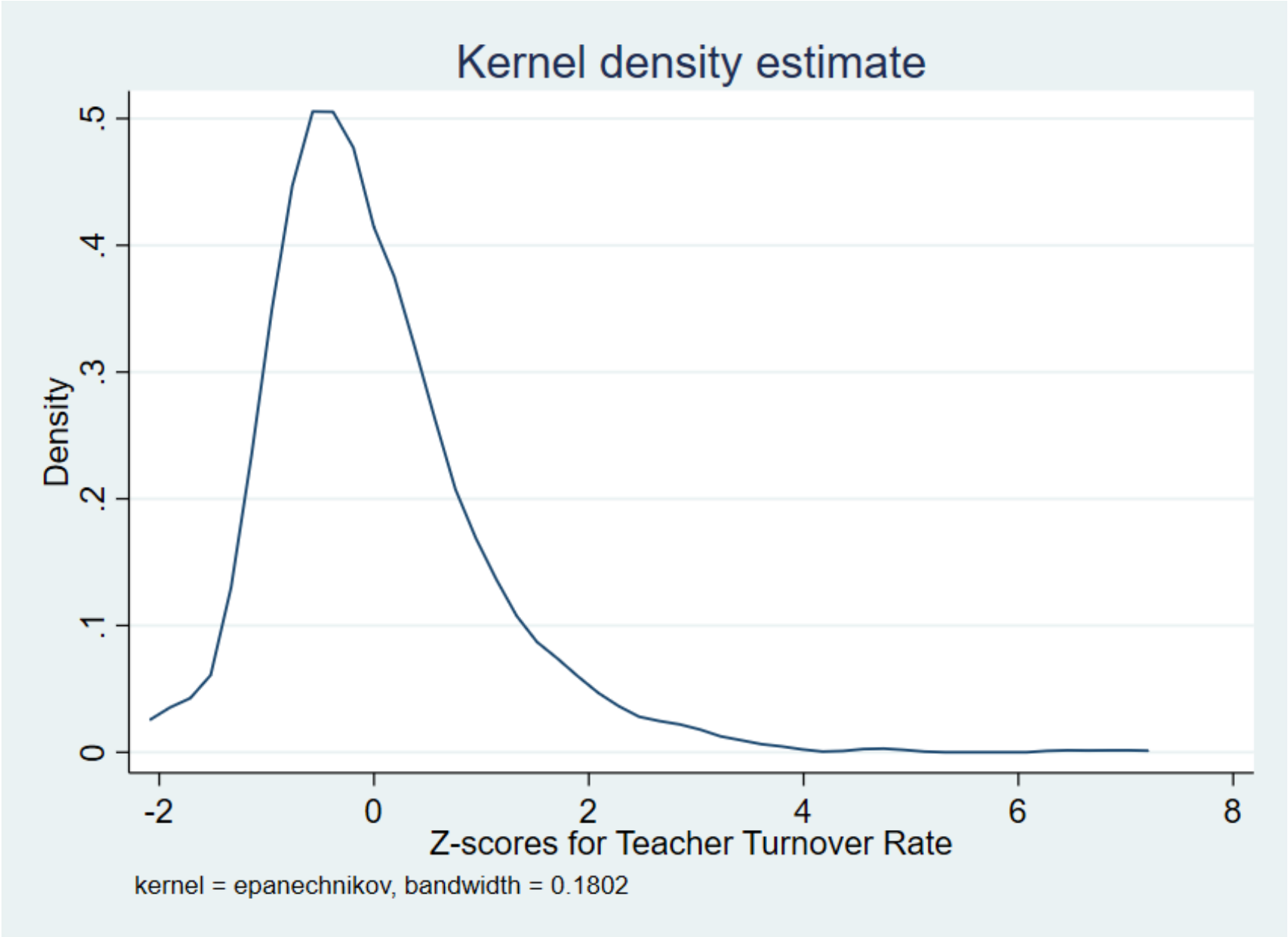
Standardization

- Sometimes it's useful to convert a variable in a dataset into a different scale so that it has a mean of 0 and a standard deviation of 1 (just like the standard normal distribution)
- Standardization: converting a variable so that its values indicate how many standard deviation-units each data point is from the mean of the distribution (creating a distribution with a mean of 0 and a standard deviation of 1)
- Z-score: how we refer to a variable (or data point) that has been standardized

Standardization: Texas school districts



Standardization: Texas school districts



Standardization

$$z = \frac{x - \mu}{\sigma}$$

- Standardizing doesn't change the shape of the distribution (similar to re-labeling the x axis on a histogram)
- Subtracting the mean moves the distribution horizontally so that it's centered at zero
- Dividing by the standard deviation stretches or squishes the distribution so that the standard deviation is equal to one

Standardization

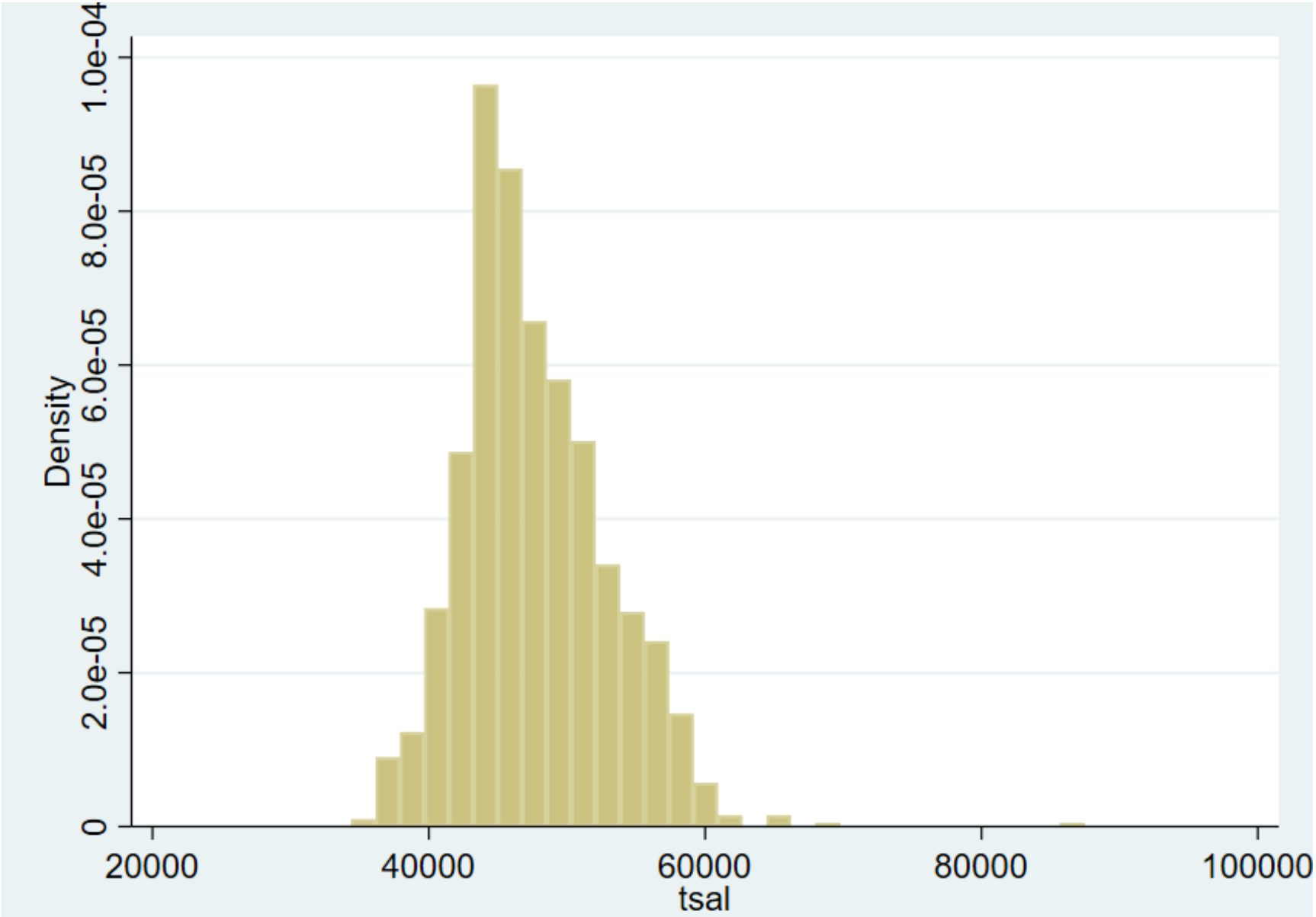
- Z scores allow us to make comparisons among distributions with similar shapes but different means/standard deviations
- Example: comparing two students if one took the ACT and the other took the SAT (could also use percentiles)

Log transformation

- Using a logarithmic transformation can help to make outliers less pronounced if most of the outliers are on the right side of the distribution (are extremely large values); log transformation won't help with outliers that are on the left side of the distribution
- While logging a variable may help with outliers, logged variables are typically harder to interpret
- Example: salary data often has outliers on the right side of the distribution

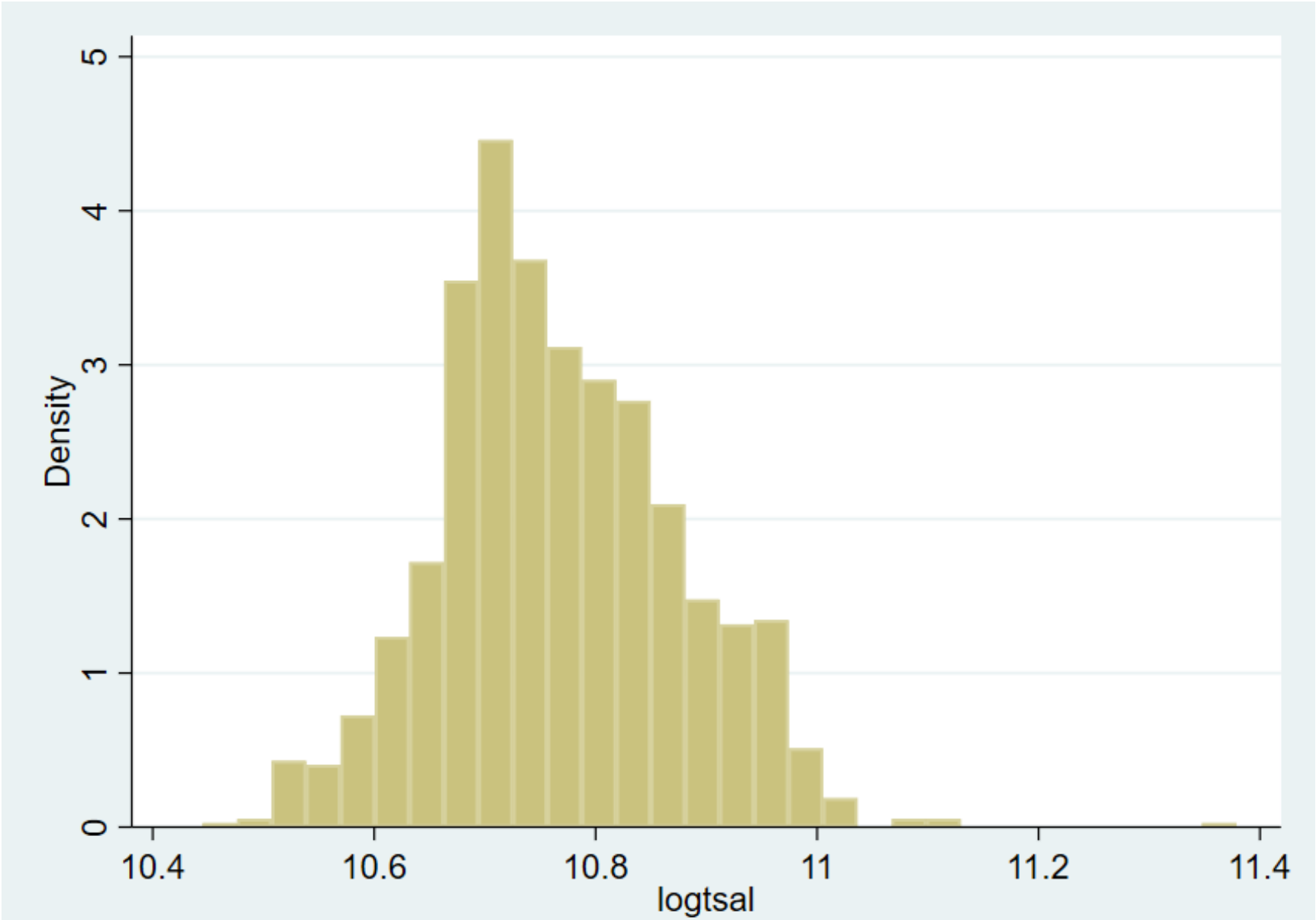
Log transformation

Average teacher salaries (before log transformation)



Log transformation

Logged average teacher salaries



Log transformation

Average teacher salaries (before log transformation) – compare mean to median

```
. sum tsal, detail

                            tsal
-------------------------------------------------------------
      Percentiles      Smallest
 1%         37300          34376
 5%         40144          36091
10%         41748          36391       Obs                1,196
25%       44057.5          36760       Sum of Wgt.        1,196

50%         46801                      Mean           47653.11
                        Largest        Std. Dev.      5251.786
75%         50924          66004
90%         55099          66176       Variance       2.76e+07
95%         57102          68072       Skewness       .8008082
99%         60203          87448       Kurtosis       5.517889
```

Log transformation

Logged teacher salaries – note that the mean and median are now closer

```
. sum logtsal, detail

                            logtsal
-------------------------------------------------------------
      Percentiles      Smallest
 1%     10.52675       10.44511
 5%     10.60023        10.4938
10%     10.63941       10.50208      Obs                 1,196
25%     10.69325       10.51217      Sum of Wgt.         1,196

50%     10.75366                     Mean              10.76584
                       Largest       Std. Dev.         .1076784
75%     10.83809       11.09747
90%     10.91689       11.10007      Variance          .0115946
95%     10.95259       11.12832      Skewness          .3394967
99%     11.00548        11.3788      Kurtosis          3.538327
```

Log transformation

- Log transformations can also sometimes be useful if we want to examine things in percentage terms (e.g., the Black-White wage gap example in Wheelan)
- To convert from a log-change to a percentage-change (approximated), calculate the change in the log values, and then move the decimal place over two places to the right to get the percentage (just as when converting a proportion to a percentage)
  - Suppose a teacher used to have a log salary of 10.5 but now it is 10.6. That is a .1 (or a .10) increase. Moving over the decimal point two places and adding a percentage sign gives us 10%.
  - Similarly, a teacher making a log-salary of 10.9 makes about 10% more than someone making 10.8
- Feel free to try out the math for this on your own with a couple examples