Social Statistics

Week Three, Class One

September 26, 2023

```r
library(tidyverse)
library(palmerpenguins)
penguins |>                                                                    ①
  mutate(                                                                      ②
    bill_ratio = bill_depth_mm / bill_length_mm,
    bill_area  = bill_depth_mm * bill_length_mm
  )
```

Assignment 1 General Thoughts

- Include your qmd file if you email me with questions
- Remember to add your name and date to the header
- Render as you go so it’s easier to identify where problems are
- Load packages when you load your data. And when loading tidyverse, do not include echoes, warnings, and messages
- Review in-class notebooks before starting

Assignment 1 Recap

1. What are the mean and median of `agekdbrn`?

```r
summary(assignment1$agekdbrn)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   12.0    20.0    24.0    24.7    29.0    57.0 
```

This also works…

```r
median(assignment1$agekdbrn)
```

```
[1] 24
```

```r
mean(assignment1$agekdbrn)
```

```
[1] 24.70305
```

Assignment 1 Recap

2. Find the 33rd and 67th percentiles:

```r
quantile(assignment1$agekdbrn, c(.33, .67))
```

```
33% 67% 
 21  27 
```

Assignment 1 Recap

3. What is the mode of `agekdbrn` for respondents who completed 12 or fewer years of education?

```
table(assignment1$agekdbrn
      [assignment1$educ<=12])
```

```
 12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29  30  31 
  3   8  18  55 152 235 372 389 391 395 247 256 203 216 134 119  94  62 115  41 
 32  33  34  35  36  37  38  39  40  41  42  43  45  46  47  48  50  52 
 59  41  29  26  13  15  16  11  10   4   3   2   3   2   4   1   1   1 
```

Assignment 1 Recap

Want to sort?

```
sort(
      table(assignment1$agekdbrn
      [assignment1$educ<=12]), # add comma here
      decreasing = TRUE # to sort from highest to lowest
      )
```

```
 21  20  19  18  23  22  17  25  24  16  26  27  30  28  29  32  15  31  33  34 
395 391 389 372 256 247 235 216 203 152 134 119 115  94  62  59  55  41  41  29 
 35  14  38  37  36  39  40  13  41  47  12  42  45  43  46  48  50  52 
 26  18  16  15  13  11  10   8   4   4   3   3   3   2   2   1   1   1 
```

Assignment 1 Recap

4. What proportion of respondents completed exactly 16 years of education?

```r
prop.table(table(assignment1$educ))
```

```
         8          9         10         11         12         13         14
0.01782136 0.02203944 0.03163556 0.04291891 0.28060740 0.07740167 0.13592745
        15         16         17         18         19         20
0.04871876 0.17958452 0.03975535 0.06200569 0.02214489 0.03943900
```

Want to round?

```r
round(prop.table(table(assignment1$educ)),3) # 3 for 3 decimal places
```

```
    8     9    10    11    12    13    14    15    16    17    18    19    20
0.018 0.022 0.032 0.043 0.281 0.077 0.136 0.049 0.180 0.040 0.062 0.022 0.039
```

Assignment 1 Recap

5. Use tidyverse functions to create a new data frame with only the agekdbrn and educ variables, and that is limited to respondents who have 14 or more years of education.

```r
library(tidyverse) # load the package if necessary
```

A Couple Options…

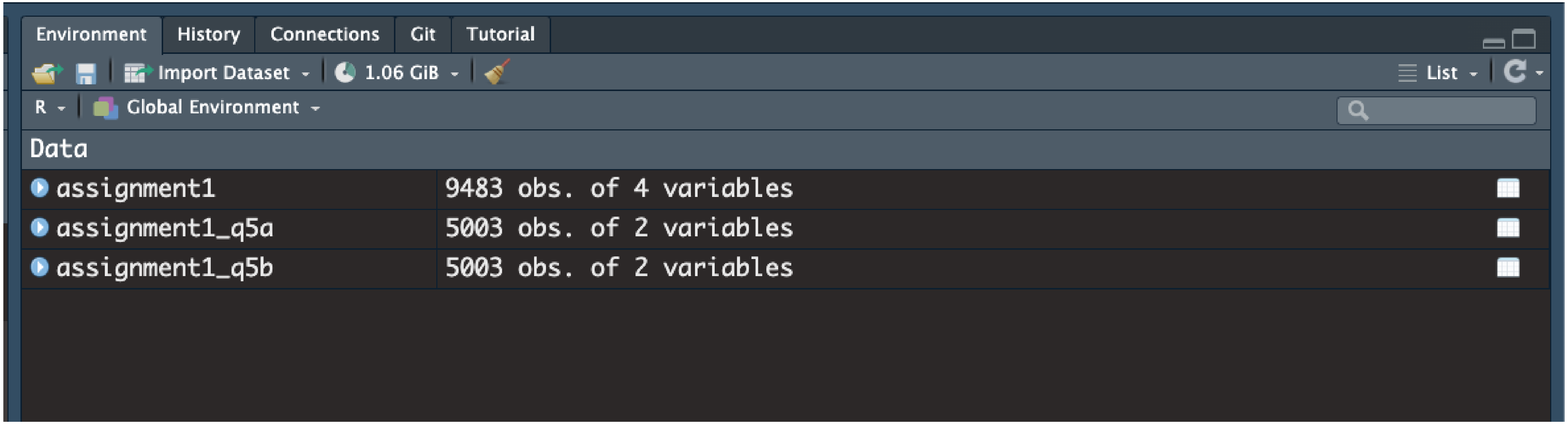
```r
# Option 1

assignment1_q5a <- select(assignment1, agekdbrn, educ) # DF name but no $
assignment1_q5a <- filter(assignment1_q5a, educ>=14) # use new DF name
```

```r
# Option 2

assignment1_q5b <- assignment1 |> # With pipe, need DF name in first line
     select(agekdbrn, educ) |> # But omit DF name from subsequent lines
     filter(educ>=14)
```

Assignment 1 Recap



Data

```
assignment1        9483 obs. of 4 variables
assignment1_q5a    5003 obs. of 2 variables
assignment1_q5b    5003 obs. of 2 variables
```

Assignment 1 Recap

6. What are the mean and median of `agekdbrn` for respondents in this new data frame?

```r
summary(assignment1_q5b$agekdbrn)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  12.00   22.00   26.00   26.65   30.00   57.00 
```

Assignment 1 Recap

7. How long did the assignment take?

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  0.670   1.500   2.000   2.038   2.750   3.000 
```

Center, Spread, Shape

- Range gives us the *minimum* and the *maximum* values
- Mean and median give us the *center* of the distribution
- Mode gives us the *most frequent* value
- Also want information about the *spread* of distributions
  - → Variance
  - → Standard Deviation
  - → Skewness

Spread

- Variance = how we measure *spread* but it has no common scale
- Standard Deviation = measure of how far observations tend to be from the mean
- Standard Deviation is the square root of the variance

$$s = \sqrt{\frac{\Sigma (y_i - \bar{y})^2}{n}}$$

How do we find the variance and standard deviation in R?

Loading Files

We'll use the `gss_week3.csv` file on Canvas. Download this file, save it, and load it as `notebook_03_01.Rmd`.

Describing Spread

Start with a summary of the `agekdbrn` variable

```r
summary(gss_week3$agekdbrn)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   9.00   20.00   23.00   24.27   28.00   57.00 
```

For variance, use `var()`:

```r
var(gss_week3$agekdbrn)
```

```
[1] 34.52177
```

For standard deviation, use `sd()`:

```
[1] 5.875523
```

Describing Spread

We can show that the standard deviation is the square root of the variance:

```r
var(gss_week3$agekdbrn) # Variance
```

```
[1] 34.52177
```

```r
sqrt(var(gss_week3$agekdbrn)) # Square Root of Variance
```

```
[1] 5.875523
```

```r
sd(gss_week3$agekdbrn) # Standard Deviation
```

```
[1] 5.875523
```

```r
sd(gss_week3$agekdbrn) ^ 2 # Standard Deviation Squared
```

```
[1] 34.52177
```

Describing Spread

Would you expect more or less variation in the distribution of completed years of education (the `educ` variable)?

```r
var(gss_week3$educ)
```

```
[1] 9.563199
```

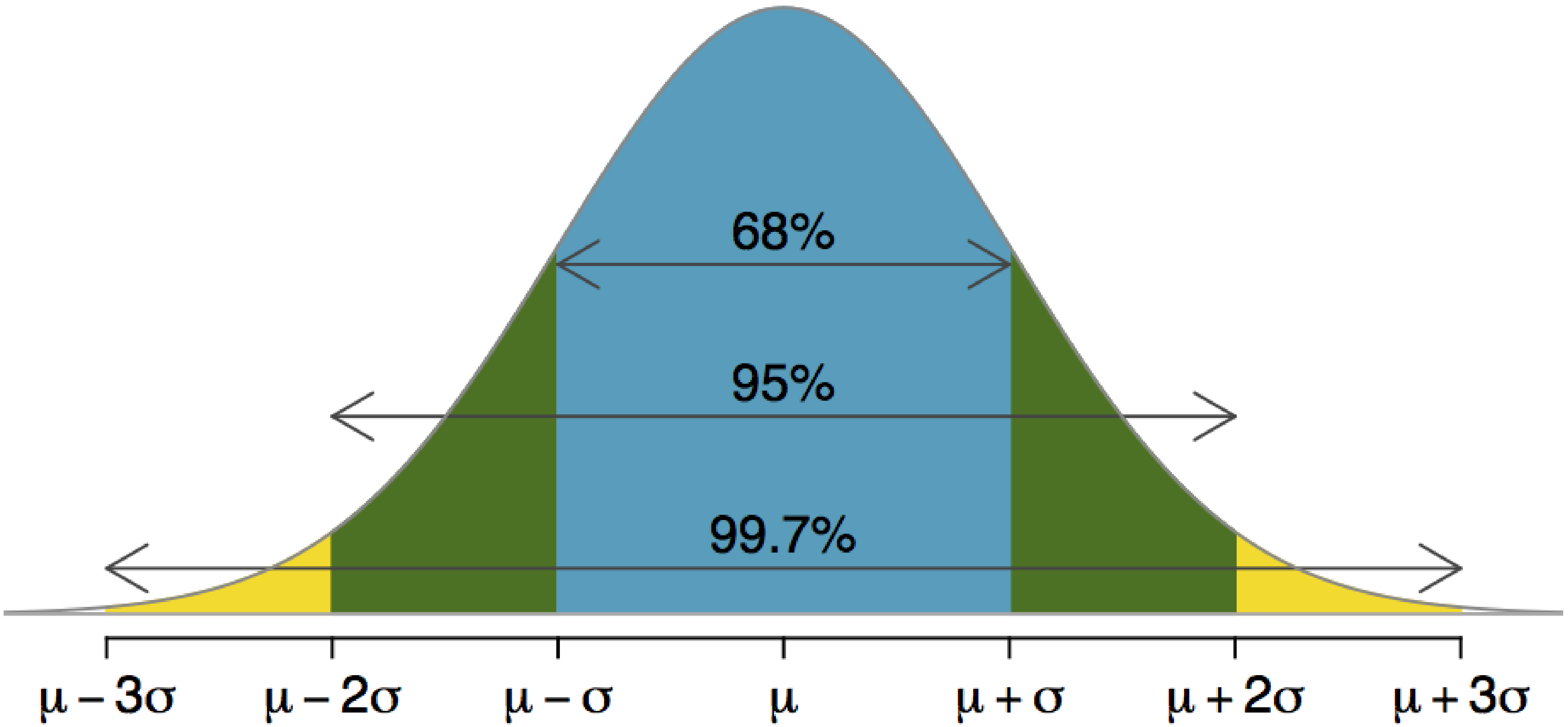
```r
sd(gss_week3$educ)
```

```
[1] 3.092442
```

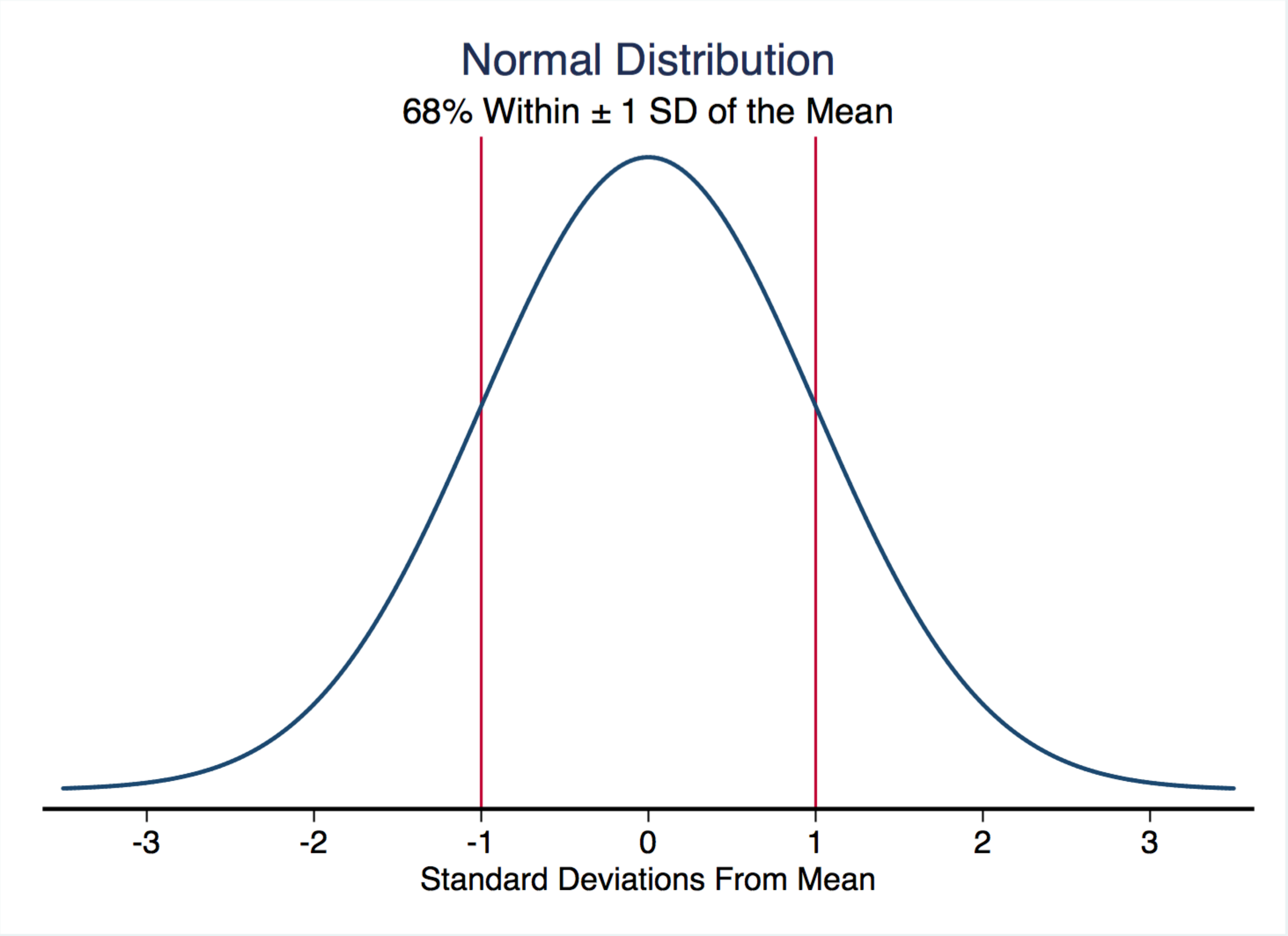
Describing The Shape of the Spread

For now, keep in mind that the shape we like the most is a normal distribution (or bell curve)

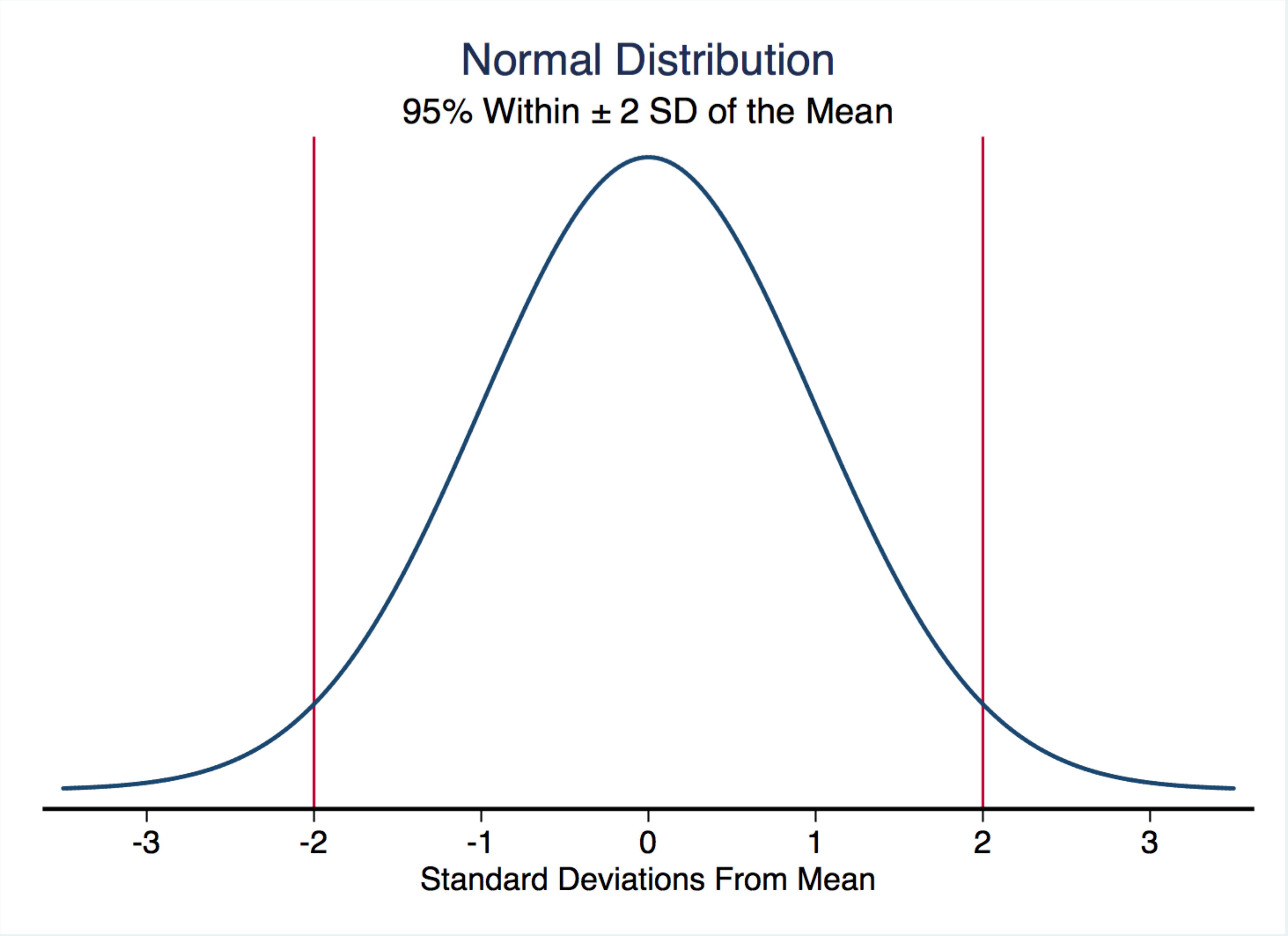
The Normal Distribution



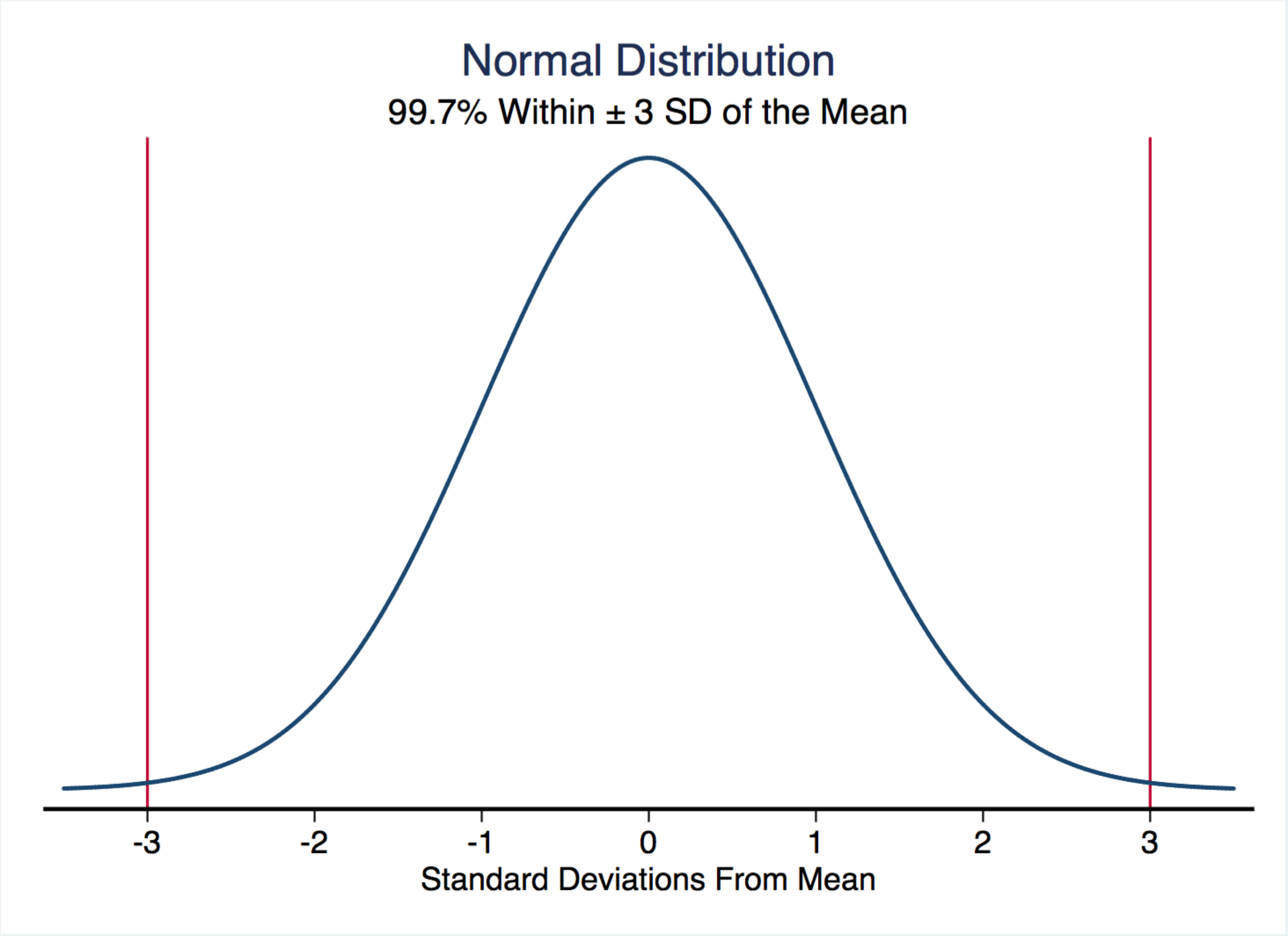
The Normal Distribution



The Normal Distribution



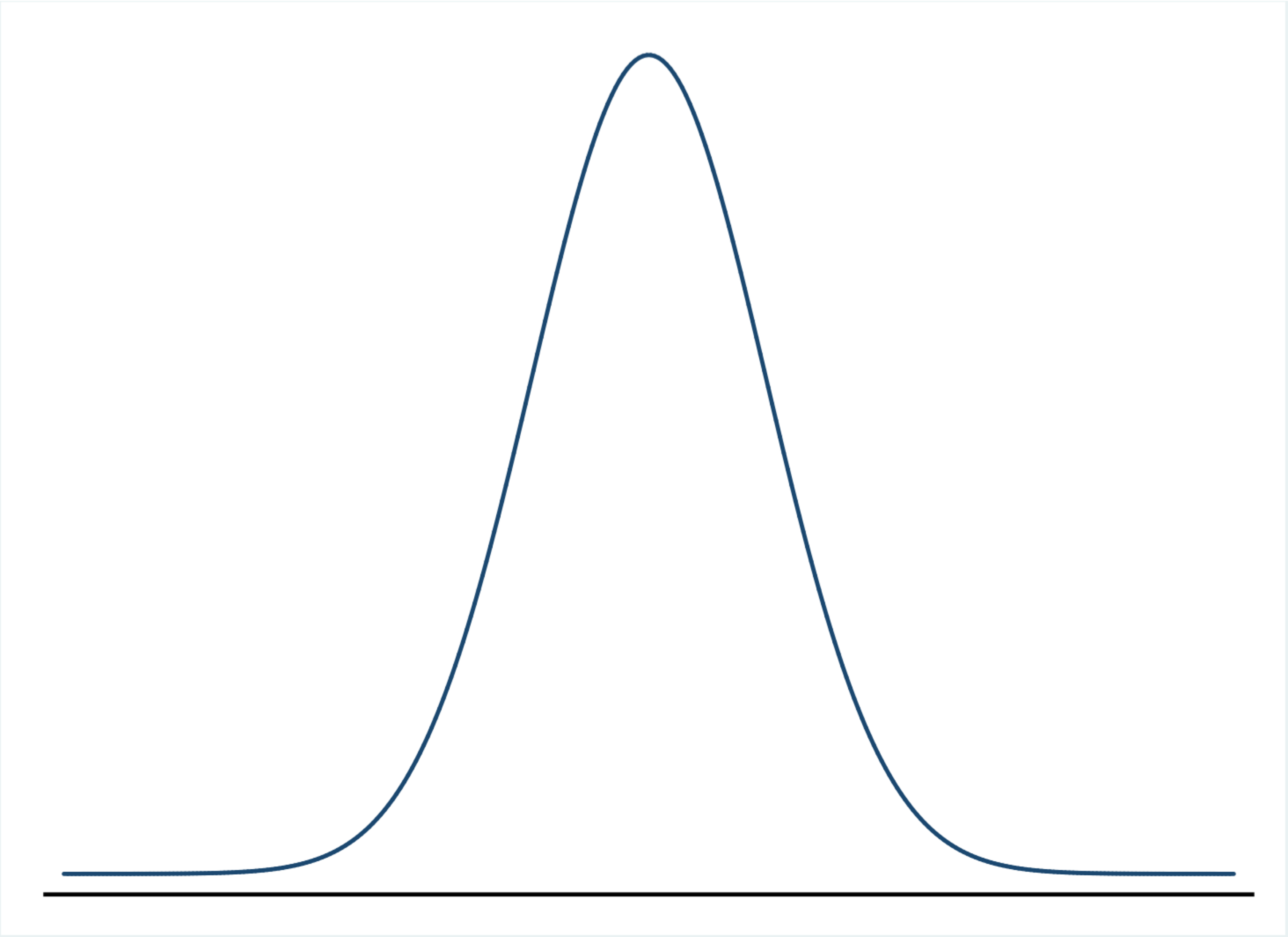
The Normal Distribution



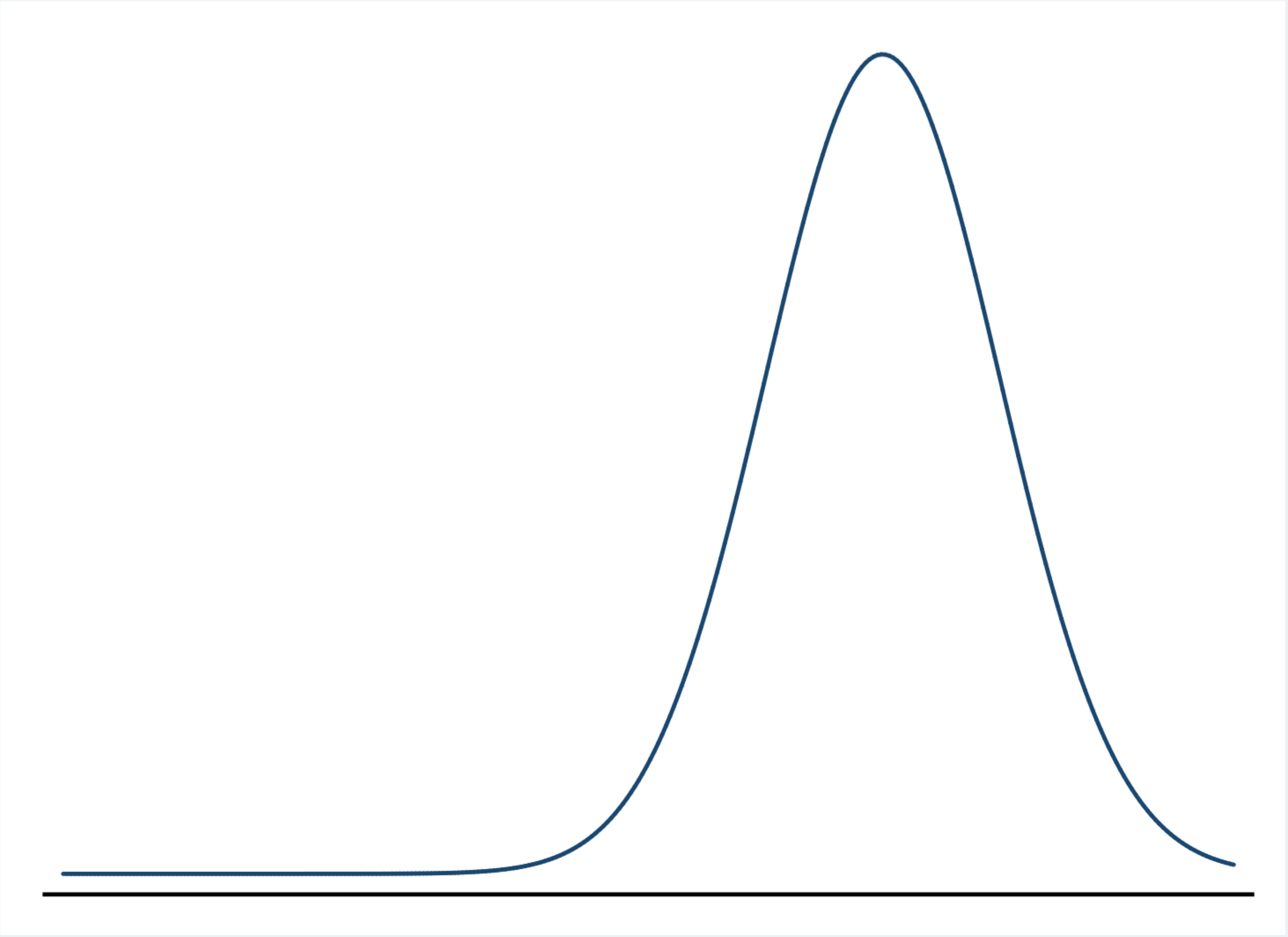
Describing The Shape of the Spread

- Since values are often not normally distributed, the measure of `skewness` tells us where the "long tail" extends
- Right skewed distributions extend to higher distributions
- Left skewed distributions extend to lower distributions

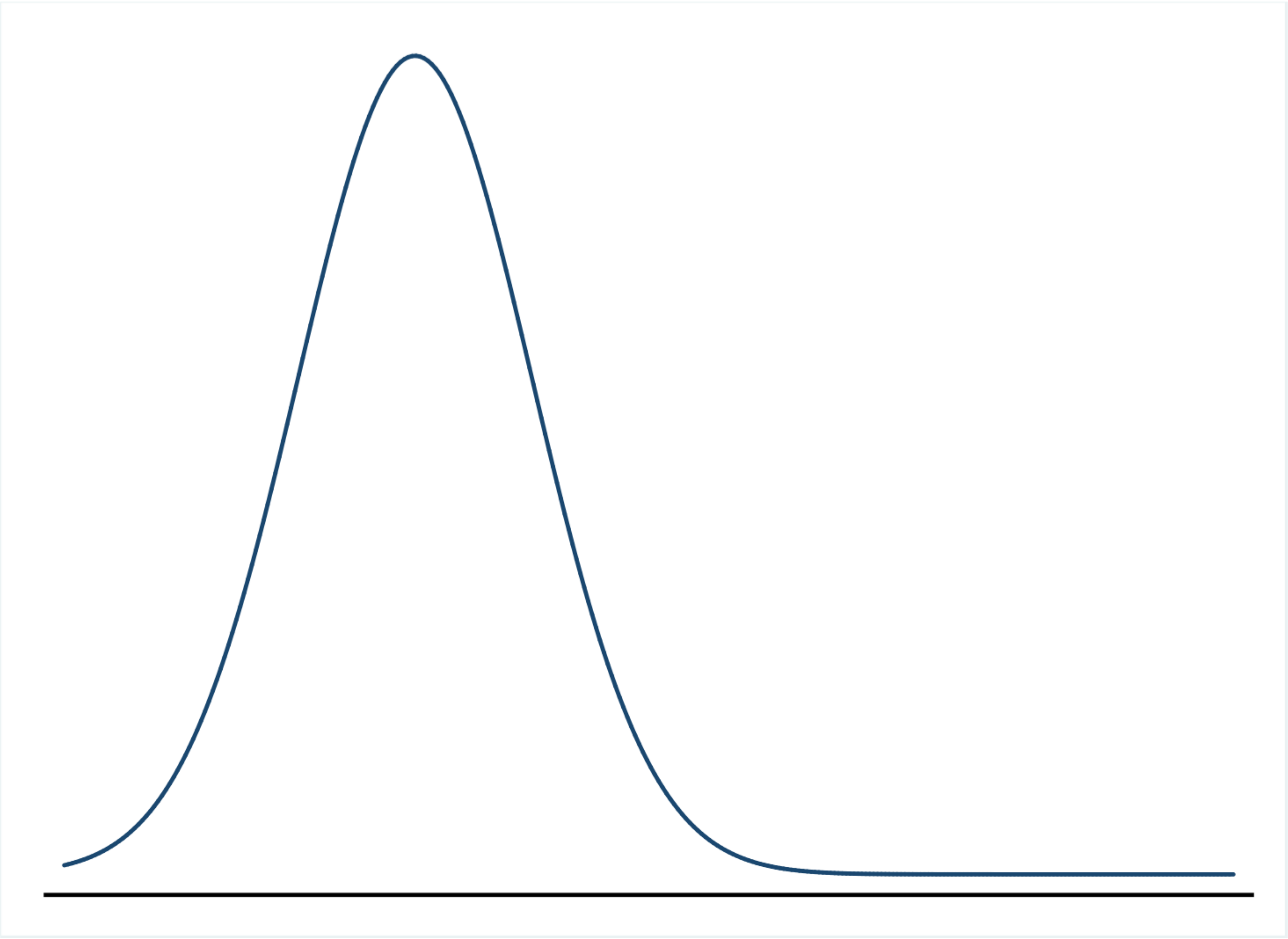
Describing Shape - Normal Distribution



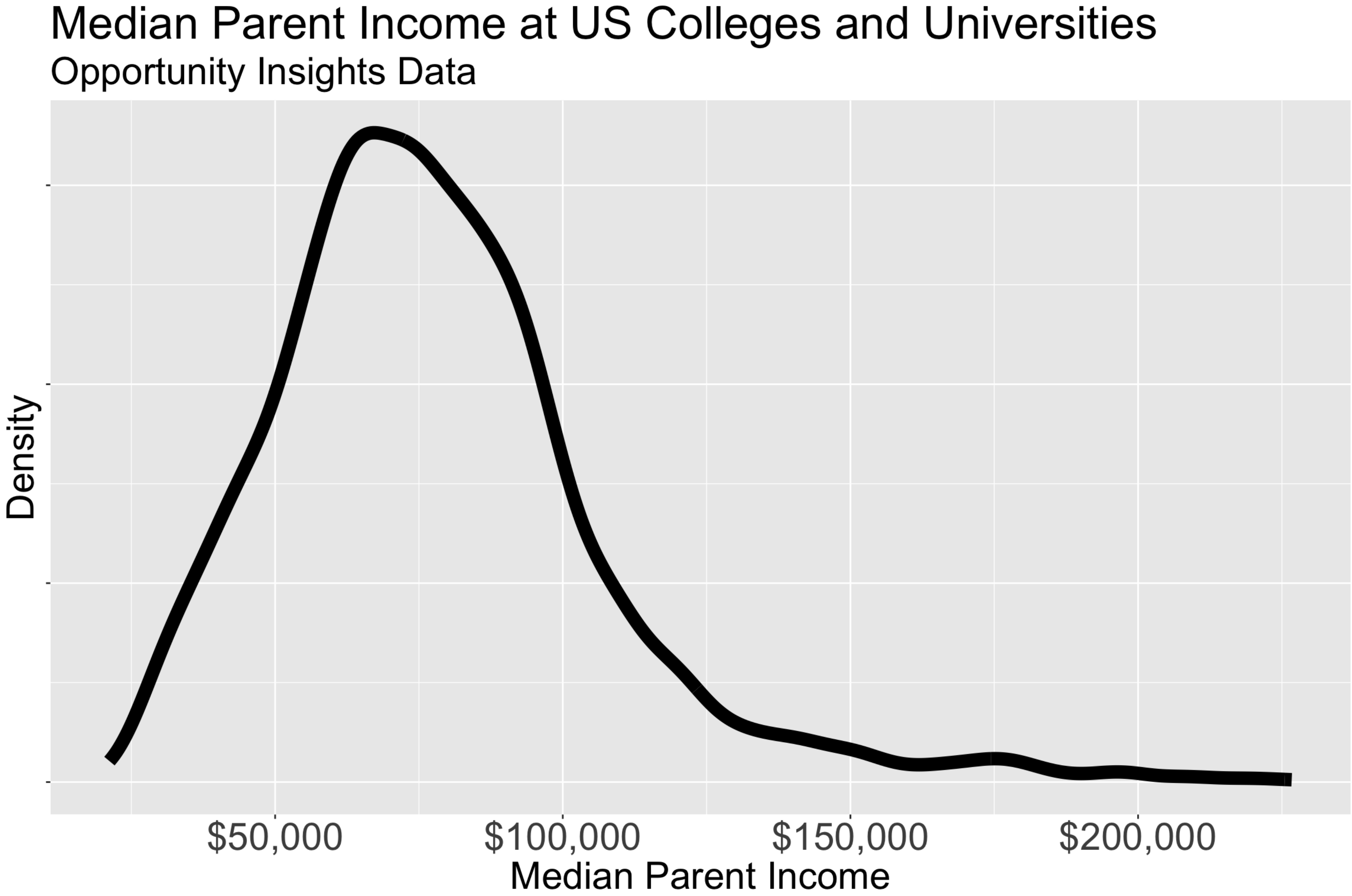
Describing Shape - Left Skew



Describing Shape - Right Skew



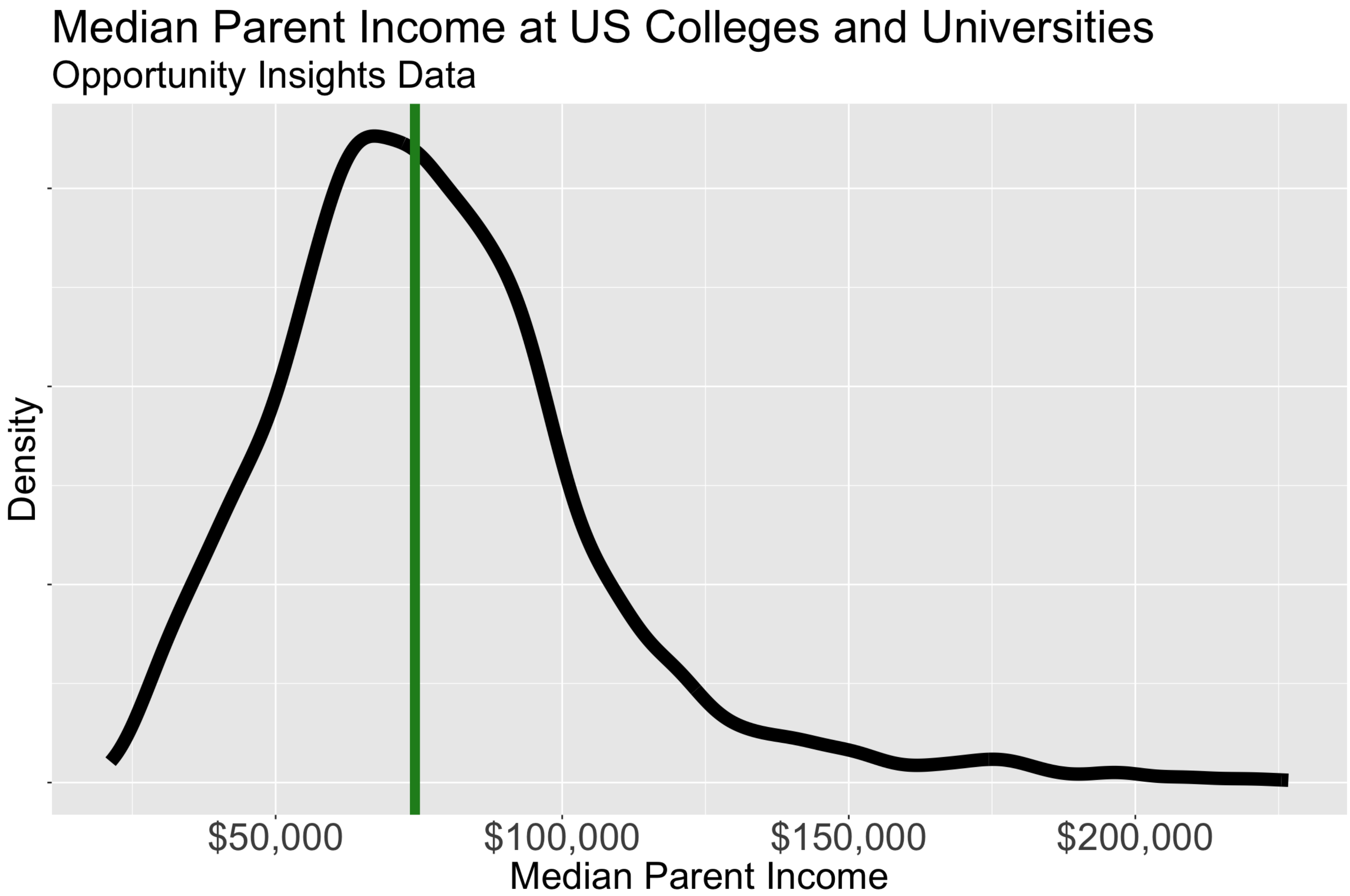
Income Is Often Right Skewed



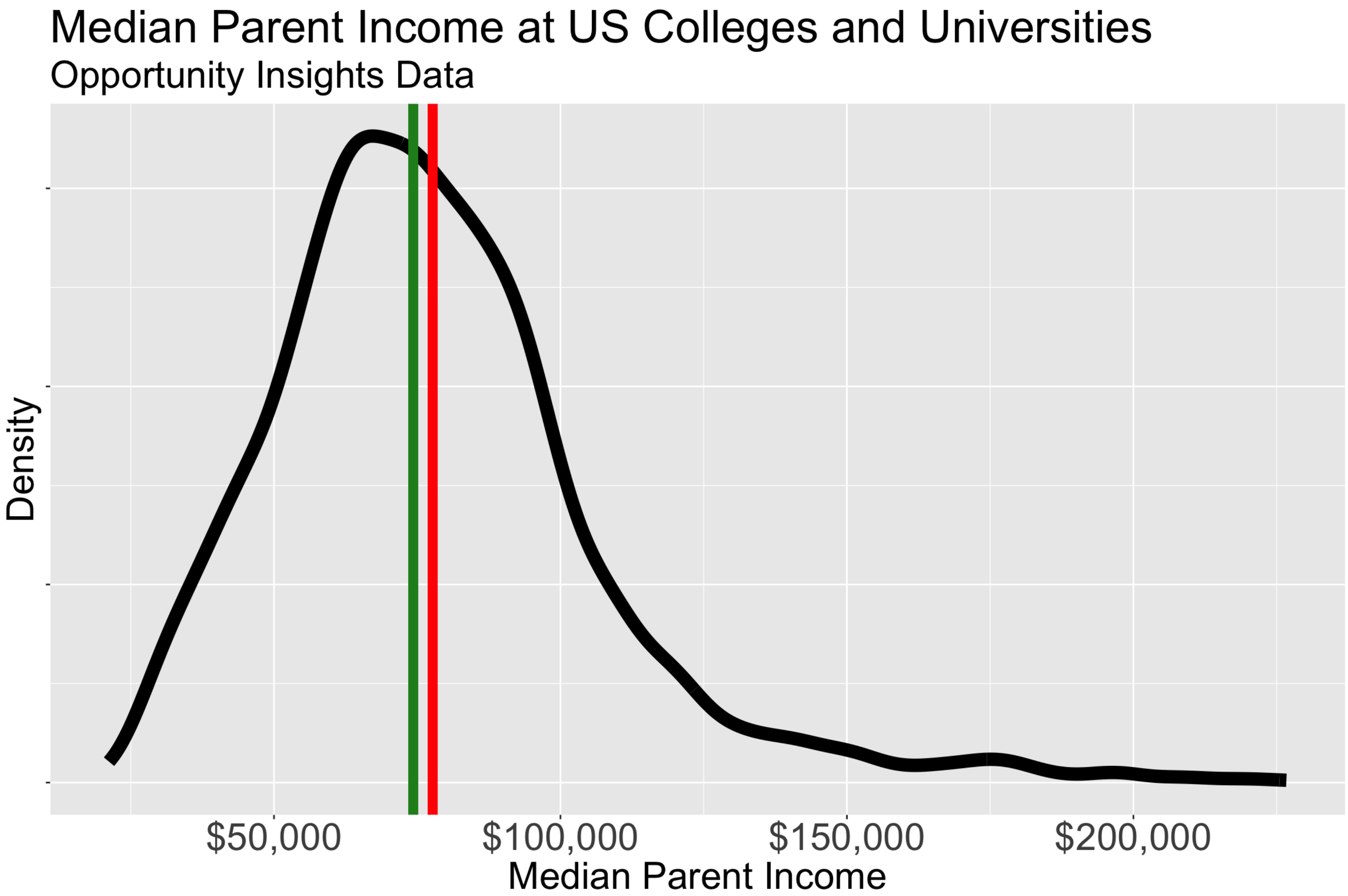
Median Parent Income at US Colleges and Universities

Opportunity Insights Data

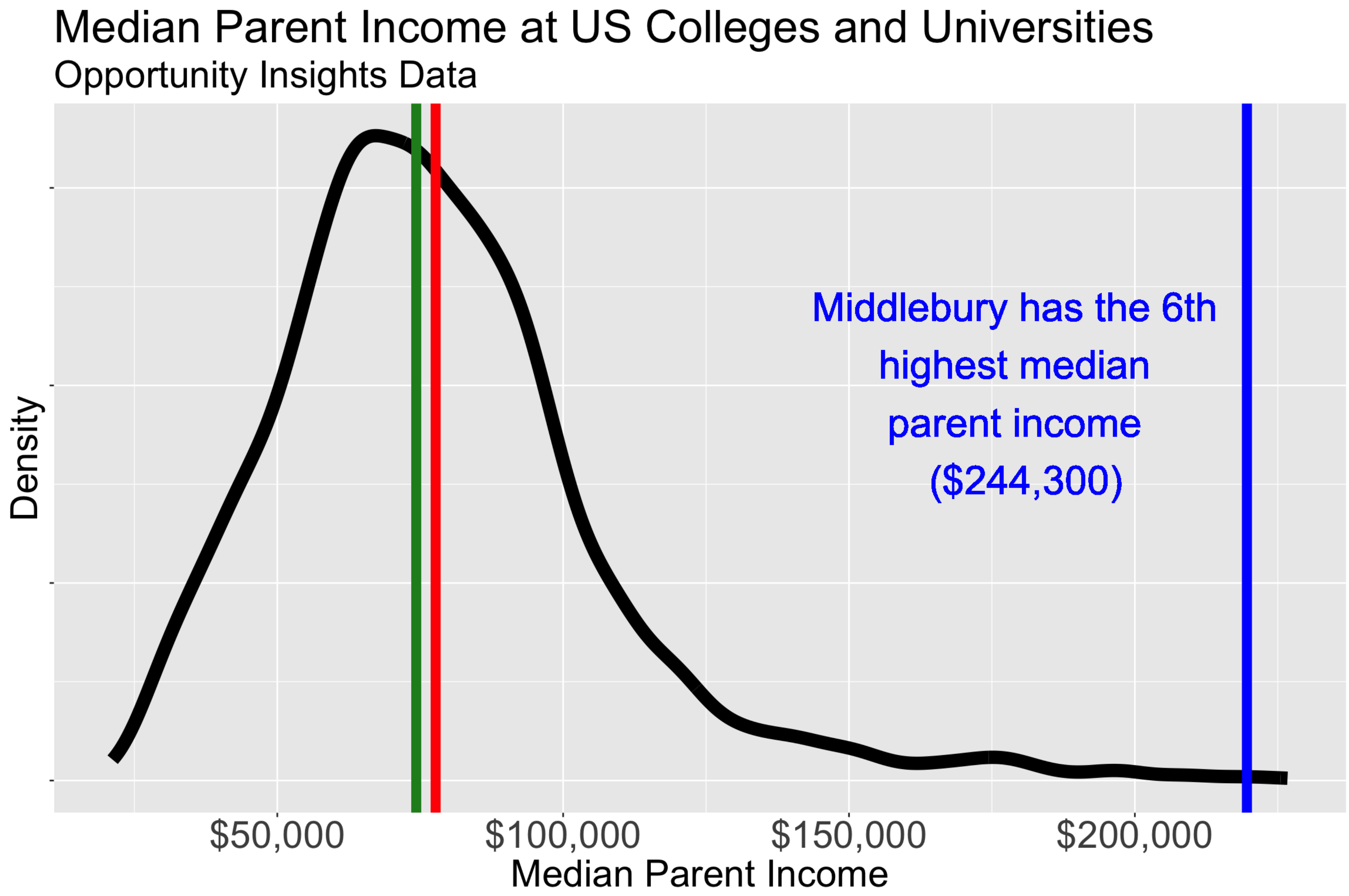
Median Not Centered



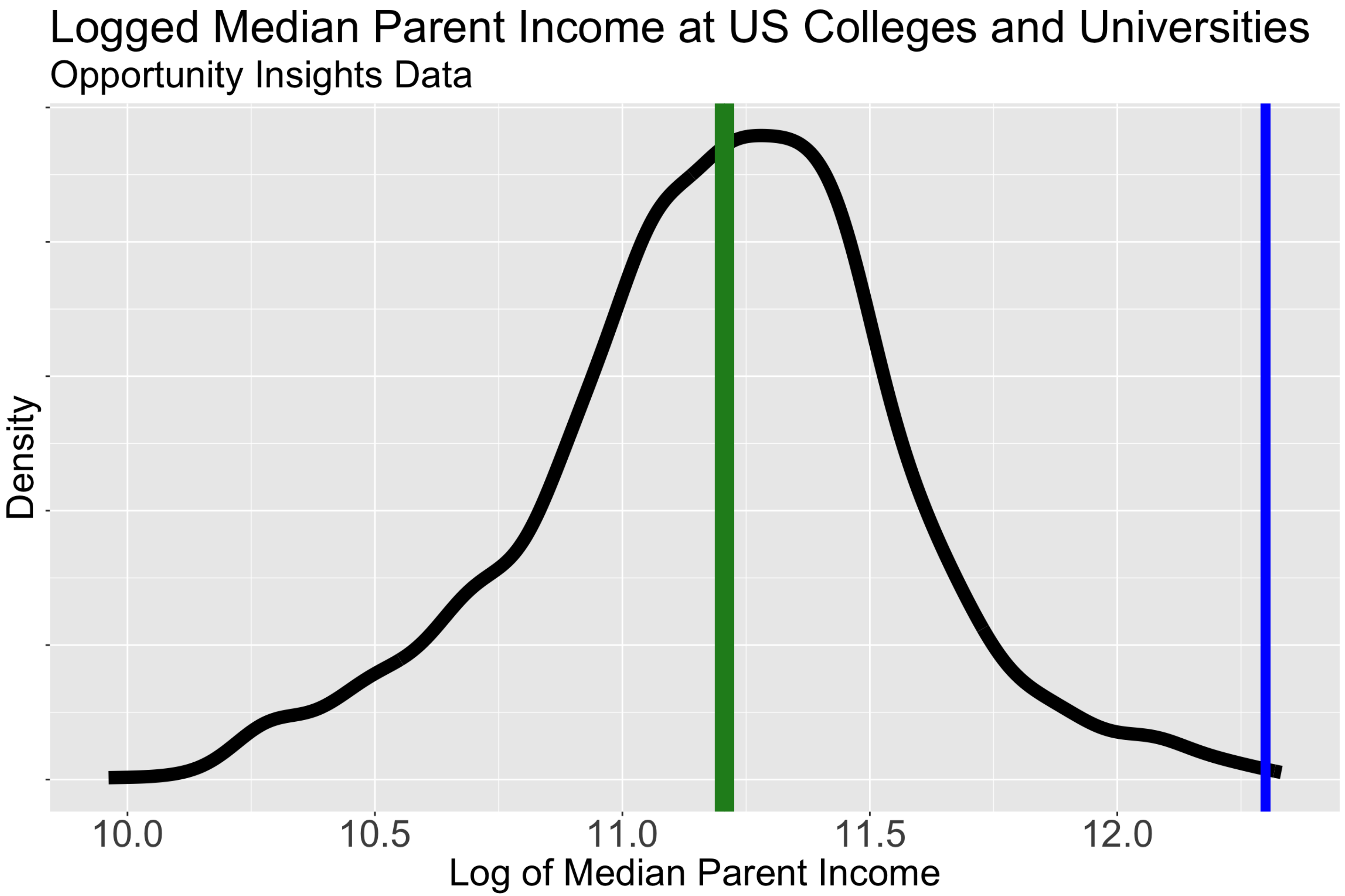
Mean Pulls To Tail



And Pulls To Highest Values



Transforming Skewed Distributions


Logged Median Parent Income at US Colleges and Universities
Opportunity Insights Data
Density
Log of Median Parent Income