Displaying data

Topic 2 · Thursday, January 7

EDUC 15: Statistics for Education Research
UCI School of Education

Where are we?

Description

Summarize data

Visualize data

We are here

Prediction

Explanation



Peter the Anteater at a UC Irvine basketball game.

Plan for today

The Mindspark study

Frequency tables

Plots and shapes

The Mindspark study

Recap: The Mindspark study

Does the Mindspark program improve children's learning?

Where

After-school centers in Delhi, India



Who

619 children, mostly in grades 6 to 9

What

Software that adapts to each child, plus small-group help from a teaching assistant

Muralidharan, Singh, and Ganimian (2019), *American Economic Review* 109(4). Photo: a Mindspark center.

A lottery split the 619 students in two

Population
Students like these

Sampling →

Sample
619 children

Random assignment →

Treatment, n = 314
Free access now

Control, n = 305
Promised the program later

Luck alone decided who got it. That makes this an experiment

Muralidharan, Singh, and Ganimian (2019).

We have scores from before and after

Baseline

Before the program

Math and Hindi tests

619 students

Endline

About 4.5 months later

Math and Hindi tests

539 students

Today we start with one column: math scores at baseline

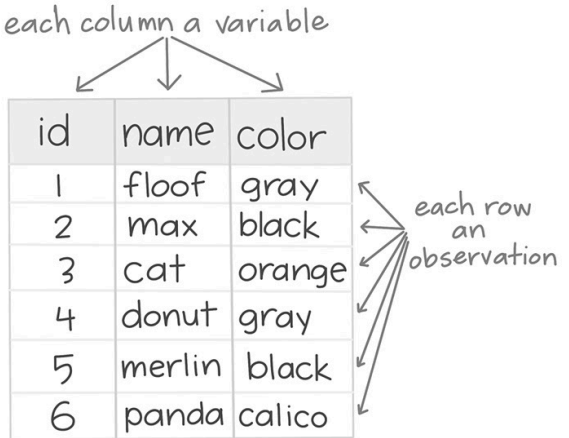
Data: Muralidharan, Singh, and Ganimian (2019).

Tidy data

> "TIDY DATA is a standard way of mapping the meaning of a dataset to its structure."
>
> –HADLEY WICKHAM

In tidy data:

- each variable forms a column
- each observation forms a row
- each cell is a single measurement



Wickham, H. (2014). Tidy Data. Journal of Statistical Software *59* (10). DOI: 10.18637/jss.v059.i10

Illustration: Lowndes and Horst (2020), Openscapes blog, CC BY 4.0. Wickham (2014), *Journal of Statistical Software* 59(10).

Tidy data: One student, one row

Student	Center	Grade	Female	Treatment	Hindi	Math
CH002	Chhatarpur	7	1	0	17	12
CH003	Chhatarpur	8	1	0	13	12
CH004	Chhatarpur	5	1	0	10	13
CH005	Chhatarpur	7	1	0	12	8
CH007	Chhatarpur	6	1	1	13	13
CH008	Chhatarpur	6	1	0	11	12

Each variable is a column. Each observation is a row. Each cell is one measurement

That's the first 6 of 619 rows. Lab 1 opens the whole file

Data: Muralidharan, Singh, and Ganimian (2019). Tidy data: Wickham (2014), *Journal of Statistical Software* 59(10).

Tidy, but where is the pattern?

12	12	13	8	13	12	12	16	10	9	10	12	12	14	12	8	13	13	20	19	7	5	17	15	10	8	13	13	11	8	10
12	8	12	13	10	5	12	14	13	12	12	9	18	10	15	12	12	4	10	14	7	15	11	11	12	15	15	10	8	10	6
5	3	14	6	15	14	10	15	8	12	14	12	13	7	13	18	10	14	6	11	13	11	13	9	11	16	14	18	12	5	16
17	19	11	14	9	9	15	15	14	14	26	2	7	11	8	11	19	15	14	14	4	12	11	15	7	15	10	20	7	9	16
9	16	13	8	8	10	9	11	13	18	9	13	10	10	6	14	16	13	12	14	16	8	8	12	12	15	14	12	10	17	6
9	11	9	9	6	11	14	13	7	11	6	14	10	13	20	6	10	21	11	10	9	14	11	7	12	8	18	11	5	5	18
5	13	9	7	14	1	3	12	15	7	14	4	17	14	2	5	5	8	5	9	12	14	12	23	20	16	12	10	5	19	9
19	12	14	16	15	10	21	16	14	17	8	12	7	9	12	9	9	8	8	13	13	14	21	24	10	15	12	3	10	5	7
11	14	12	11	10	11	12	3	8	4	13	16	10	8	10	13	10	16	6	7	3	7	13	9	12	7	12	9	10	19	12
12	21	17	8	11	4	17	12	10	10	7	9	6	16	12	11	8	6	7	18	14	11	14	10	8	11	14	10	12	10	9
8	10	7	13	11	8	10	8	16	13	9	10	9	15	15	11	5	7	13	10	11	11	10	7	12	3	10	8	12	9	11
5	9	6	12	13	16	11	9	8	21	13	11	13	12	5	10	2	10	17	15	11	6	11	13	17	14	18	11	7	16	5
17	8	10	14	13	14	15	8	8	19	10	14	12	16	7	15	14	16	10	10	15	6	5	9	12	10	11	15	9	11	9
11	9	14	13	5	14	14	14	0	11	14	7	11	15	17	10	12	7	13	12	14	11	7	13	15	14	9	10	12	12	13
18	15	17	12	14	13	9	14	10	10	9	9	16	12	2	15	11	9	4	11	10	14	7	10	18	12	14	16	11	12	5
12	9	12	9	8	10	14	11	6	11	3	12	6	8	8	11	8	9	12	17	16	4	16	19	12	8	6	7	3	14	14
15	13	10	14	13	11	15	14	19	16	11	12	7	12	12	12	7	14	13	16	14	18	10	15	8	12	13	12	9	15	8
9	12	8	7	9	3	6	5	12	4	10	8	10	11	10	13	9	9	11	10	7	14	13	12	14	10	7	13	11	9	9
15	6	13	17	18	8	14	10	6	9	8	12	13	7	10	9	4	3	10	9	13	5	14	5	14	0	15	11	8	3	8
11	4	12	11	12	5	9	11	9	12	11	9	11	7	9	12	13	11	14	9	12	11	14	10	14	10	18	12	11	10	

All 619 baseline math scores. Data: Muralidharan, Singh, and Ganimian (2019).

Predict it: The most common score

Respond at **PollEv.com/adb**

You just saw all 619 scores. The test has 35 points. Which score do you think came up most often?

A 6

B 12

C 18

D 24

Frequency tables

A frequency table counts each score

Left column: each score X

Right column: how many students, called the frequency f

Show every category, even a zero

Score X	f
20	4
21	5
22	0
23	1
24	1
25	0
26	1

Nobody scored 22 or 25

Data: Muralidharan, Singh, and Ganimian (2019).

The full table has a peak at 12

Score X	f	Score X	f	Score X	f
0	2	9	55	18	13
1	1	10	66	19	9
2	4	11	60	20	4
3	11	**12**	**76**	21	5
4	10	13	48	22	0
5	23	14	61	23	1
6	20	15	33	24	1
7	34	16	24	25	0
8	43	17	14	26	1

The f column adds up to n = 619, so every student is counted once

Data: Muralidharan, Singh, and Ganimian (2019). In the book, this total is written N.

Grouping trades detail for clarity

Ten or fewer intervals is typical

Use a simple width: 2, 5, 10 or 20

Start at a multiple of the width

Keep all intervals the same width

Interval	f
0–4	28
5–9	175
10–14	311
15–19	93
20–24	11
25–29	1

Inside 10–14, we no longer know any one student's exact score

Data: Muralidharan, Singh, and Ganimian (2019).

Try it: Group 20 real scores

11, 10, 12, 9, 8, 8, 6, 8, 11, 12, 8, 16, 4, 9, 10, 10, 12, 13, 14, 10

Make a grouped table with intervals of 5, starting at 0. Then check your total

Work alone for two minutes. Compare with a neighbor. Then we vote: how many fall in 5–9?

Data: Muralidharan, Singh, and Ganimian (2019). A random draw of 20 students.

Try it: Answer

Interval	f
0–4	1
5–9	7
10–14	11
15–19	1

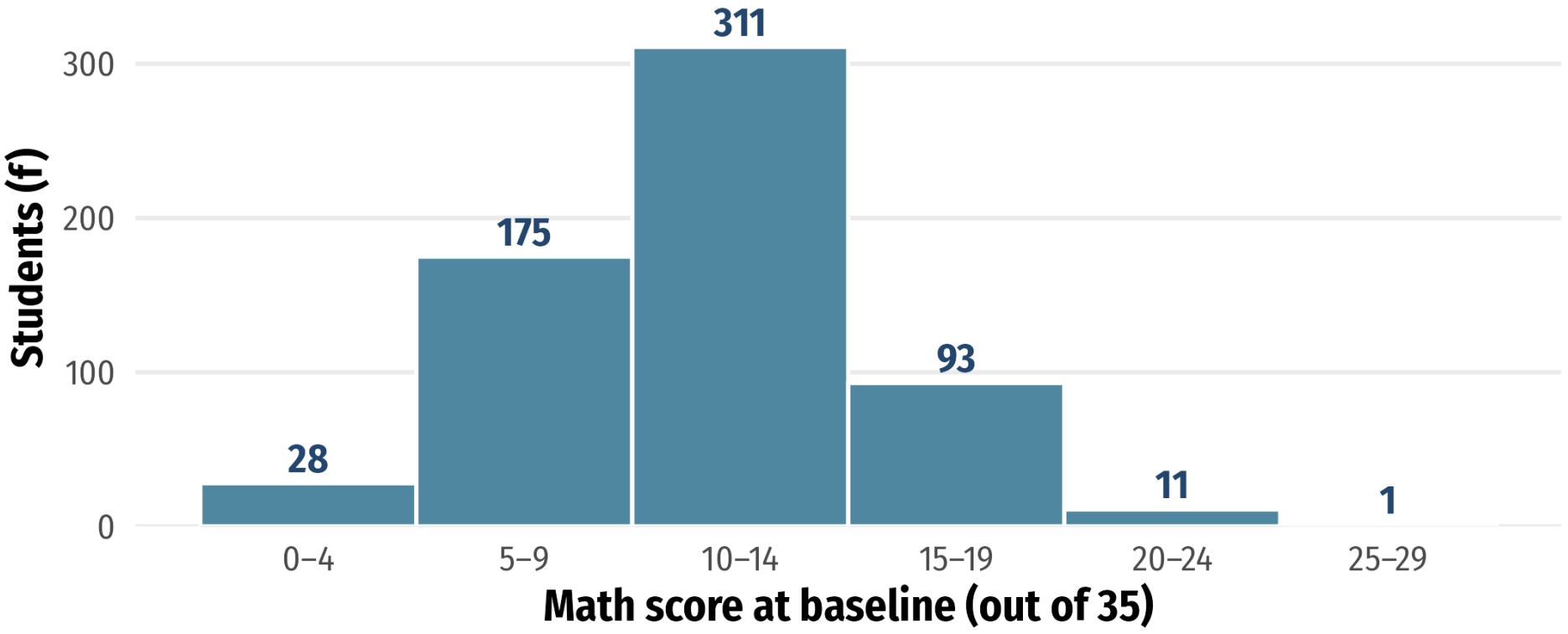
The total is 20, one count per student. Nobody got lost

Even these 20 peak at 10–14, the same interval as all 619

Data: Muralidharan, Singh, and Ganimian (2019).

Plots and shapes

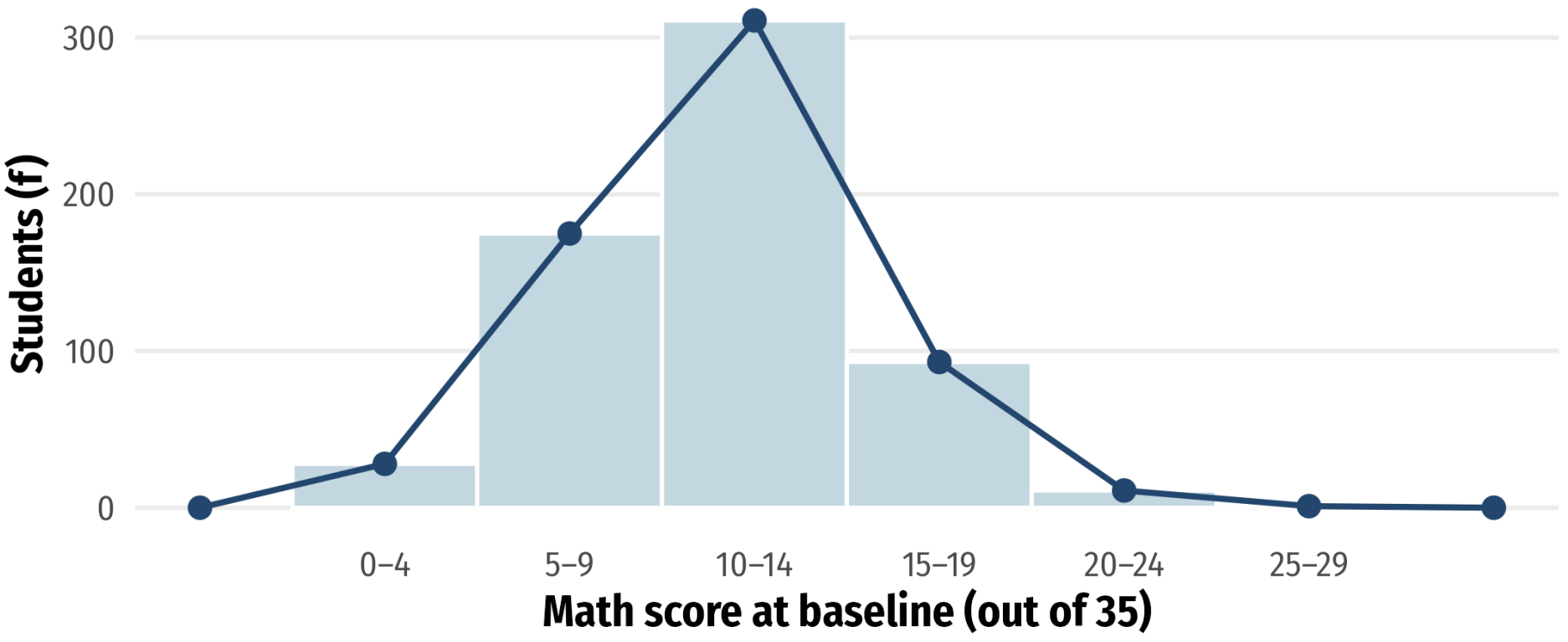
A histogram draws the grouped table



Each bar is one row of the table. The bars touch: the scale has no gaps

Data: Muralidharan, Singh, and Ganimian (2019).

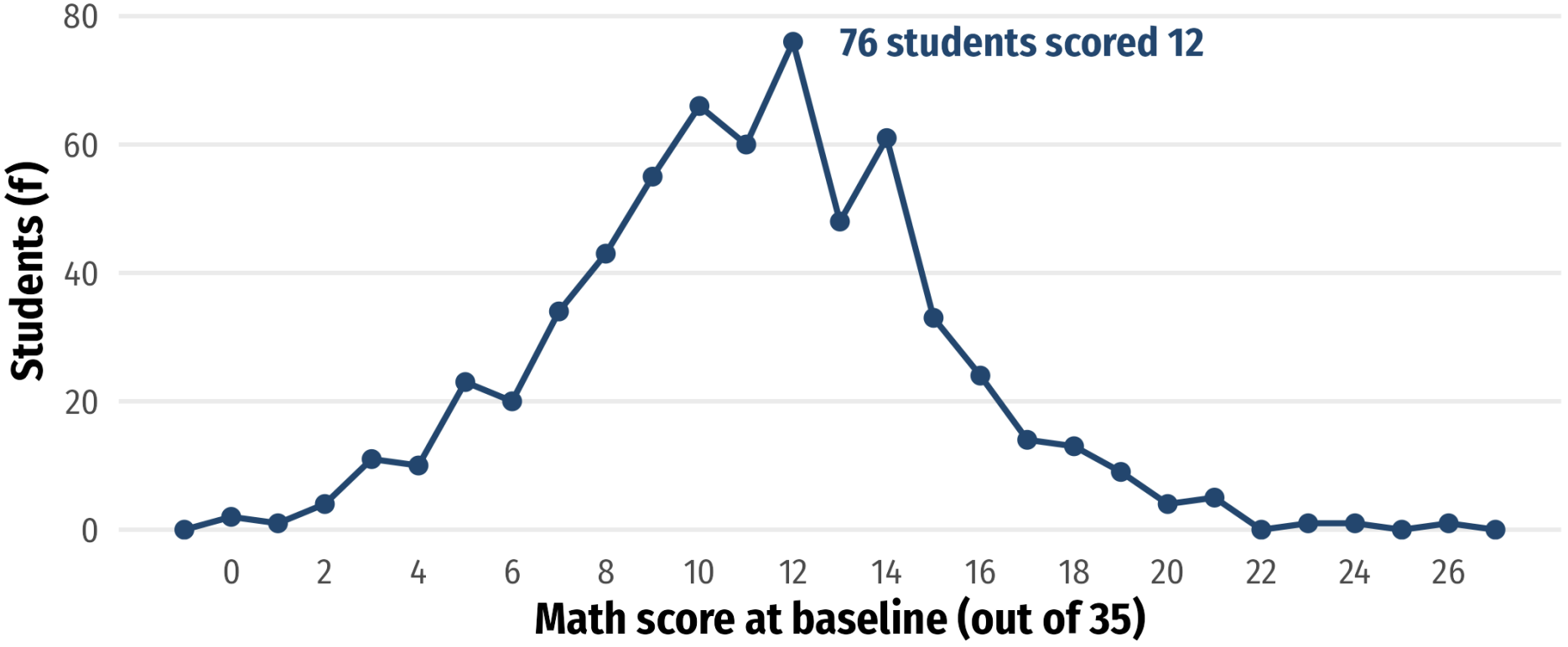
A polygon swaps bars for dots and lines



Each dot sits at the middle of its interval. A line is easier to overlay than bars

Data: Muralidharan, Singh, and Ganimian (2019).

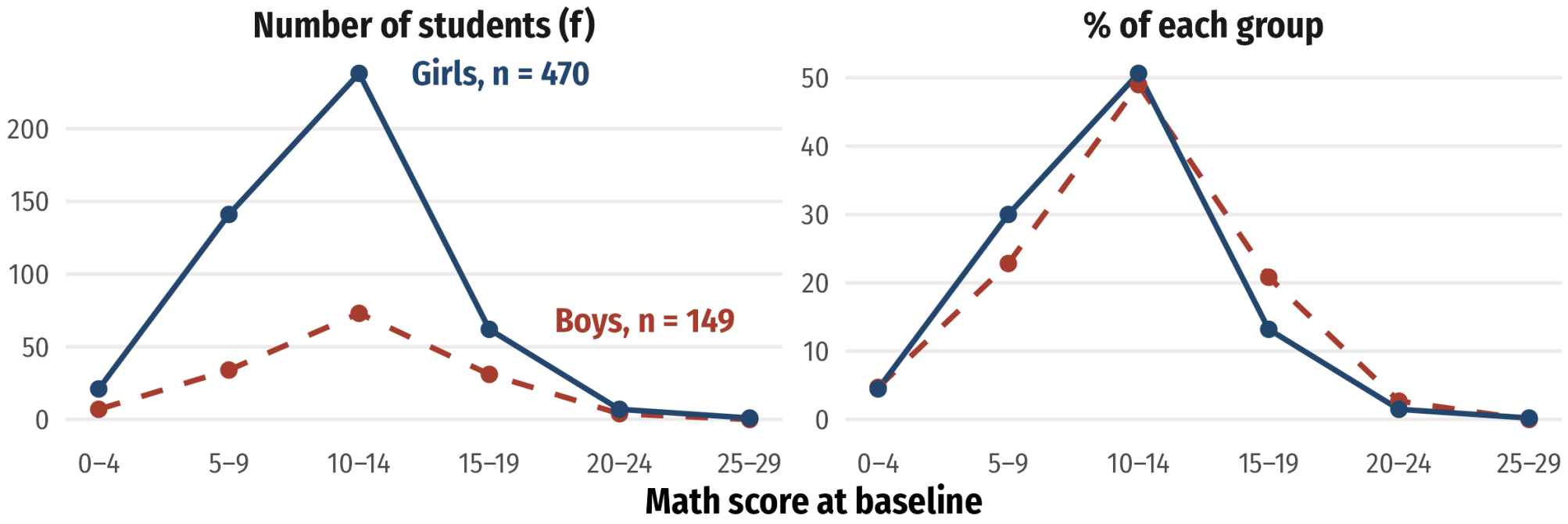
Every score: more detail, more bumps



With no grouping we keep every count, and the picture gets jagged

Data: Muralidharan, Singh, and Ganimian (2019).

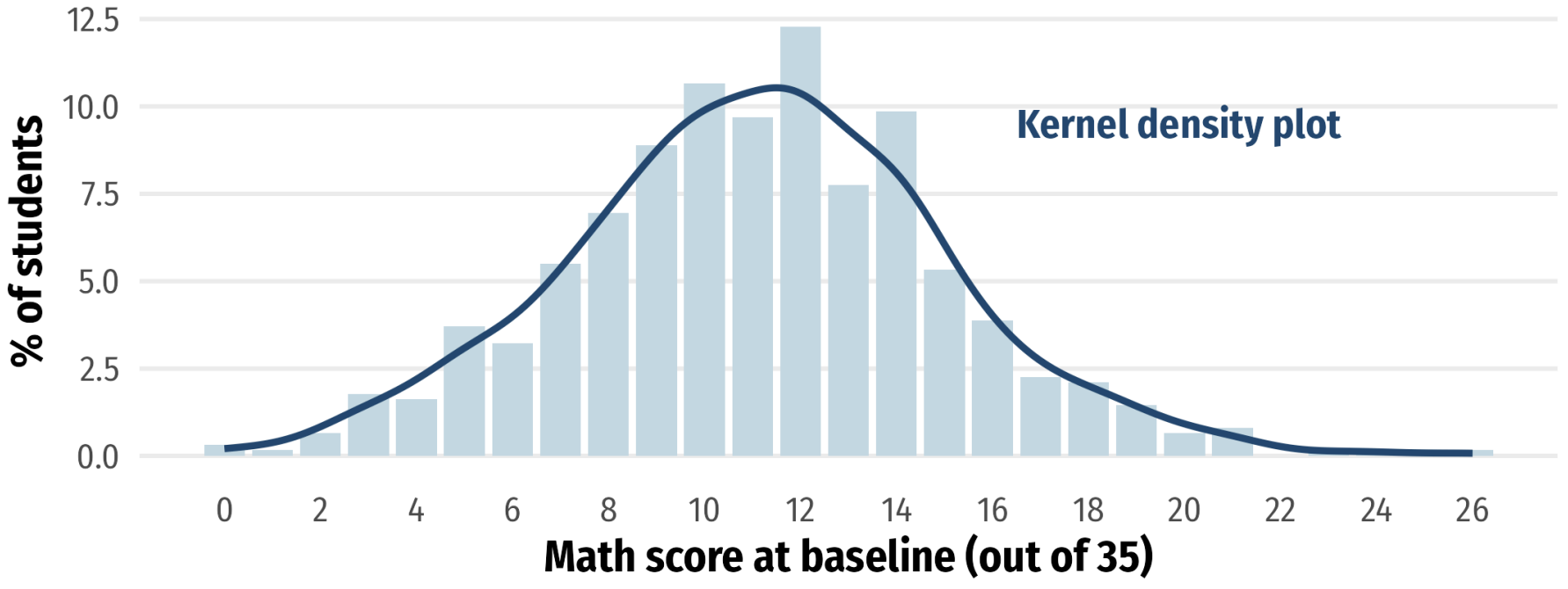
Percentages put groups on one scale



% = 100 × f / n. On the left, size hides the shape. On the right, we can compare

Data: Muralidharan, Singh, and Ganimian (2019).

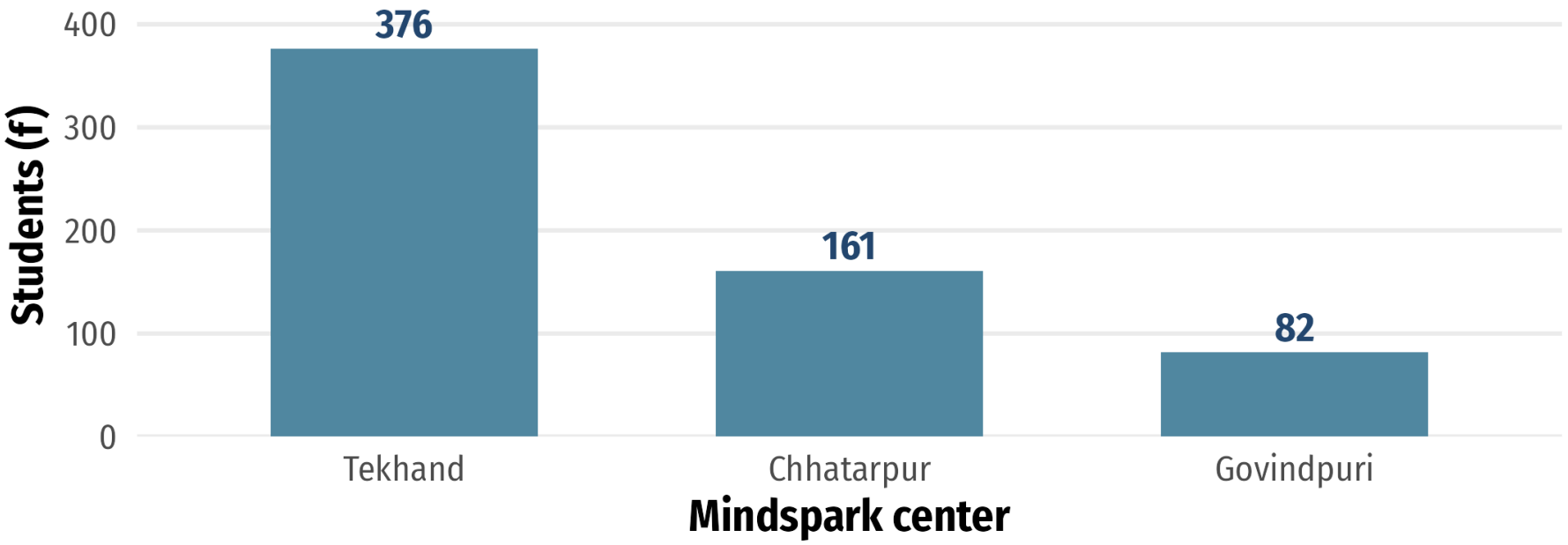
A density plot smooths out the bumps



It adds up a small bump at each score. You'll read these, not compute them

Data: Muralidharan, Singh, and Ganimian (2019).

Categories get bars with gaps

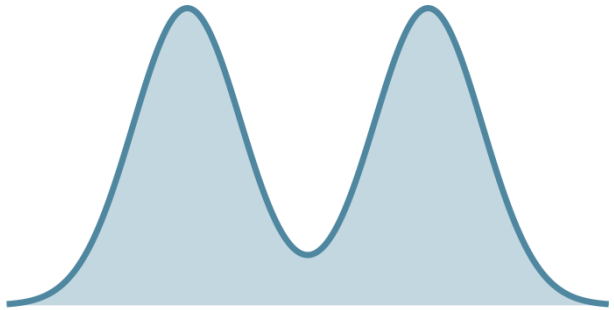


Scores have an order and even steps, so a histogram's bars touch. Centers are just names, so these bars stand apart

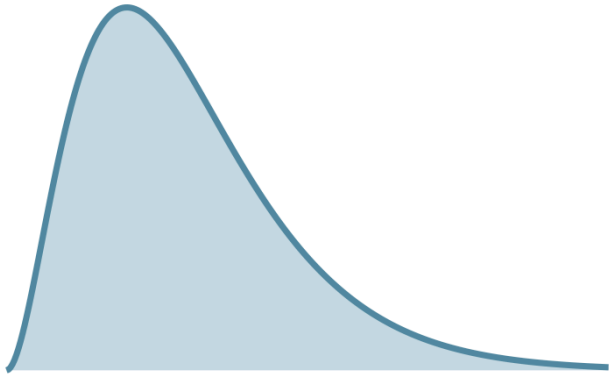
Data: Muralidharan, Singh, and Ganimian (2019).

Symmetric, unimodal and normal differ

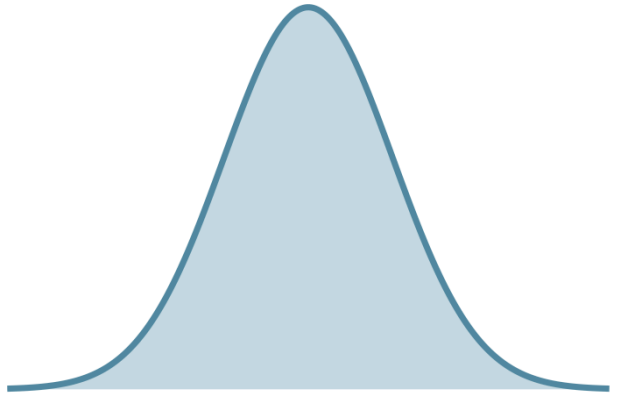
Symmetric, not unimodal



Unimodal, not symmetric



Normal: both, and a bell

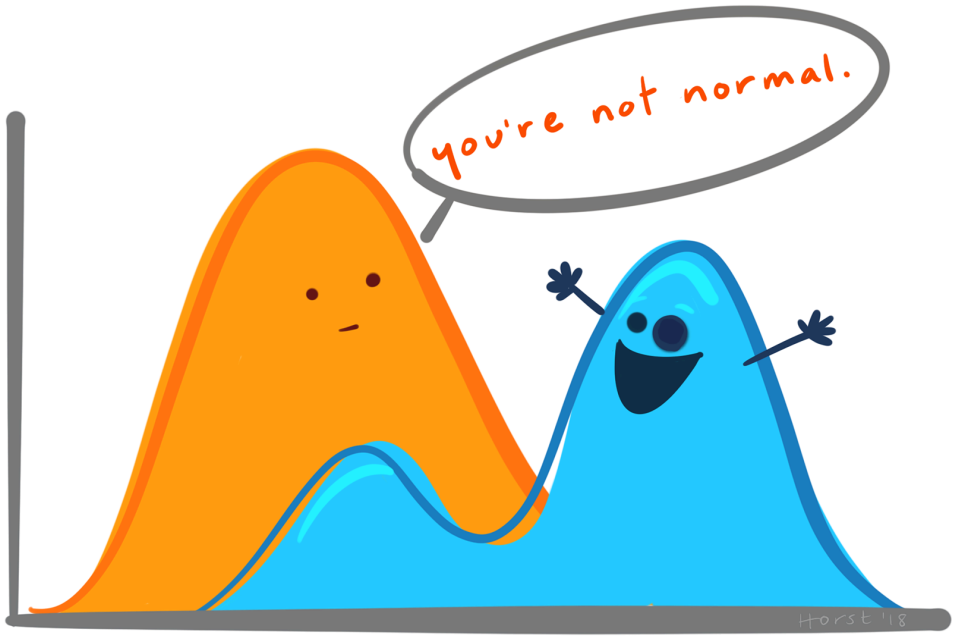


Symmetric: each side mirrors the other

Unimodal: one high point

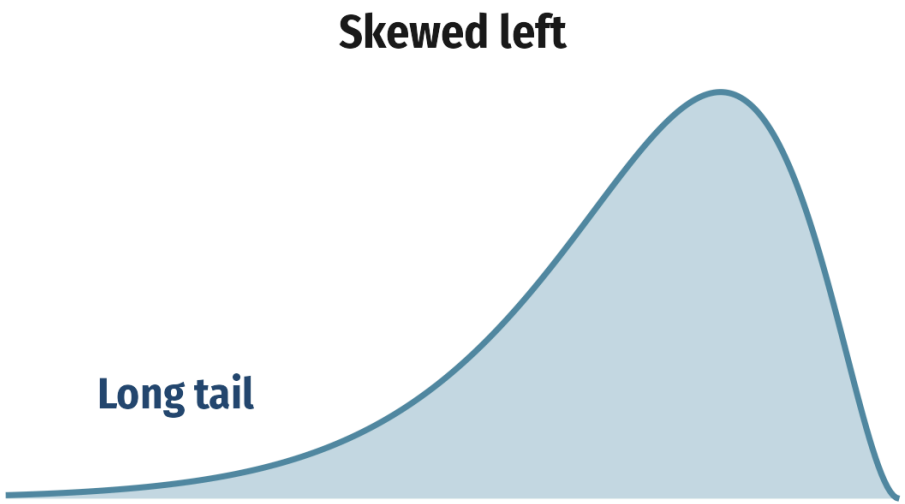
Normal: a bell, common in nature

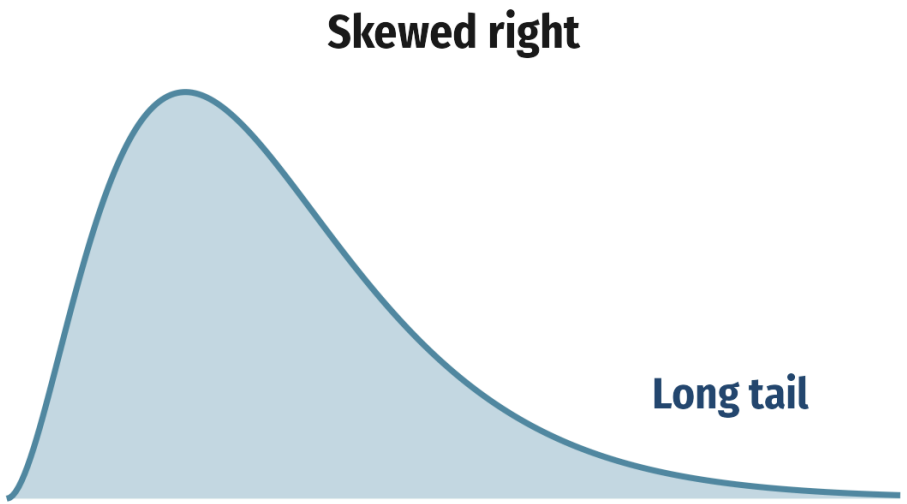
Not everything is normal



Artwork by @allison_horst, CC BY 4.0.

Skewed distributions have one long tail



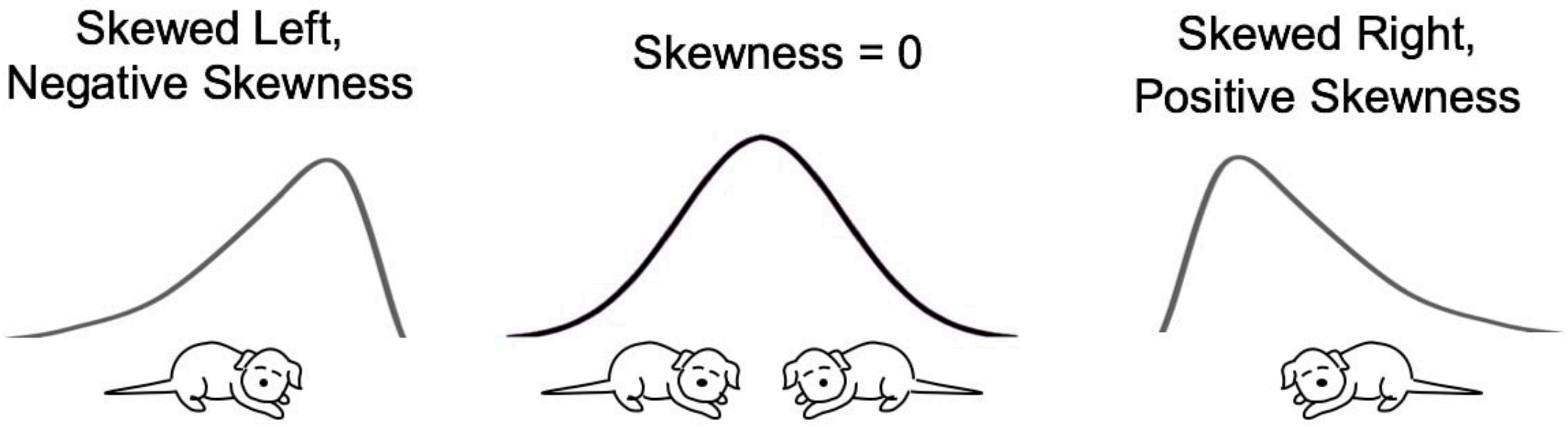


A very easy test: scores pile up high, and a few trail off to the left

A very hard test: scores pile up low, and a few trail off to the right

In the book: skewed left is negatively skewed, and skewed right is positively skewed

Everybody loves a doggo mnemonic



Reproduced by permission of John Wiley and Sons
from the book, Statistics from A to Z -- Confusing Concepts Clarified

Jawlik, *Statistics from A to Z: Confusing Concepts Clarified* (Wiley, 2016), statisticsfromatoz.com.

Predict it: The endline scores

The same study, about 4.5 months later. Picture the baseline curve you just saw. What does the endline curve look like?

A In about the same place

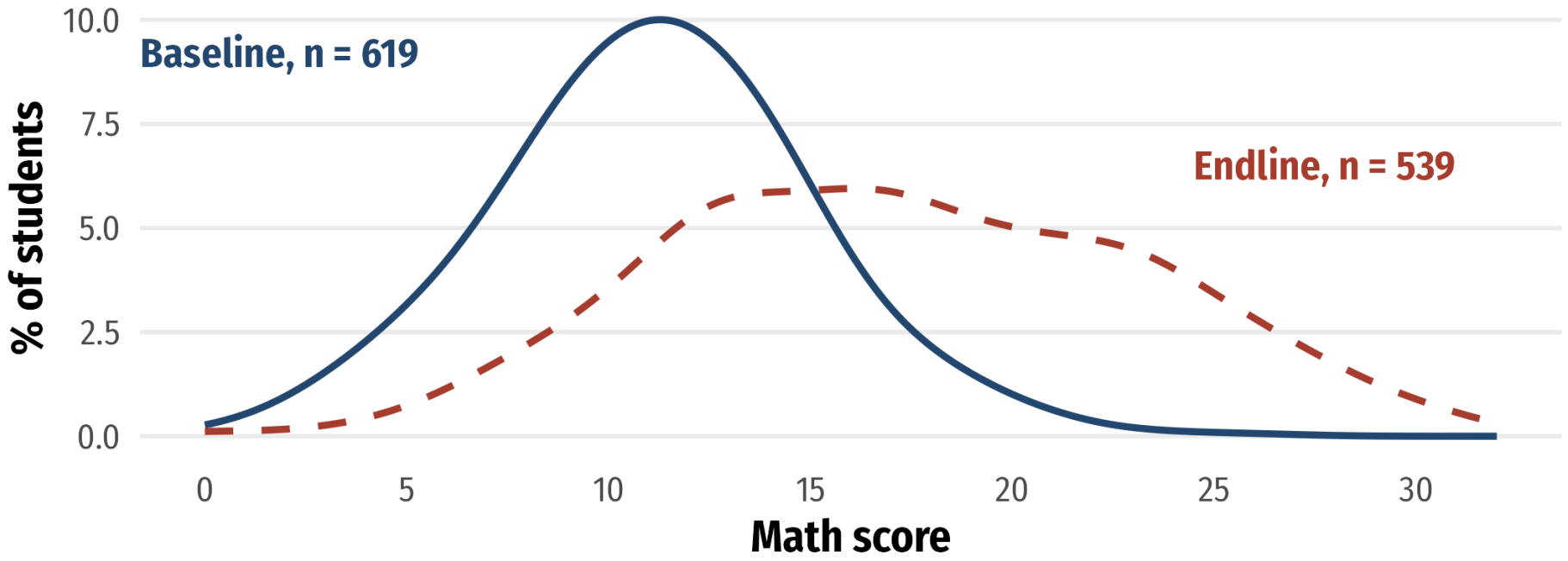
B Moved right, still one peak

C Moved right, now two peaks

D Moved left

Tell a neighbor why. Half these students had Mindspark, and half didn't

Predict it: The whole pile moved right



So, B: one wide peak. Half the students scored at least 11 at baseline, and half at least 17 at endline

Data: Muralidharan, Singh, and Ganimian (2019).

What you can do now

Say what the Mindspark study did, and why it used a lottery

Build a frequency table, ungrouped or grouped

Read a histogram, a polygon, a density plot and a bar graph

Describe a shape: symmetric, unimodal, normal or skewed

Next steps

Before Tuesday

Read Chapters 3 and 4

Tomorrow: Lab 1

Meet the data. Go to your own lab, and bring a calculator and a laptop

Nothing is due this Sunday. The first MindTap work is due January 17 at 6 PM