DSC 10, Spring 2018
Lecture 8

Histograms

sites.google.com/eng.ucsd.edu/dsc-10-spring-2018

Credit: Anindita Adhikari and John DeNero

Area Principle

What is wrong with this picture?

"New iPad Has a Gigantic 70-Percent Larger Battery"

Source: Gizmodo

What is wrong with this picture?

“New iPad Has a Gigantic 70-Percent Larger Battery”

Source: Gizmodo

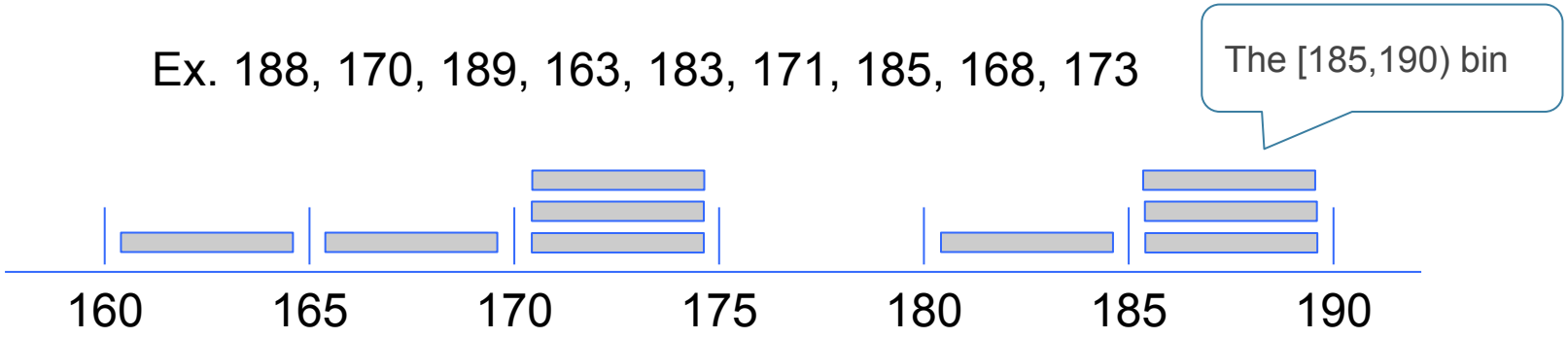
Binning

Binning Numerical Values

Binning is counting the number of numerical values that lie within ranges, called bins.

- Bins are defined by their lower bounds (inclusive)
- The upper bound is the lower bound of the next bin

Ex. 188, 170, 189, 163, 183, 171, 185, 168, 173

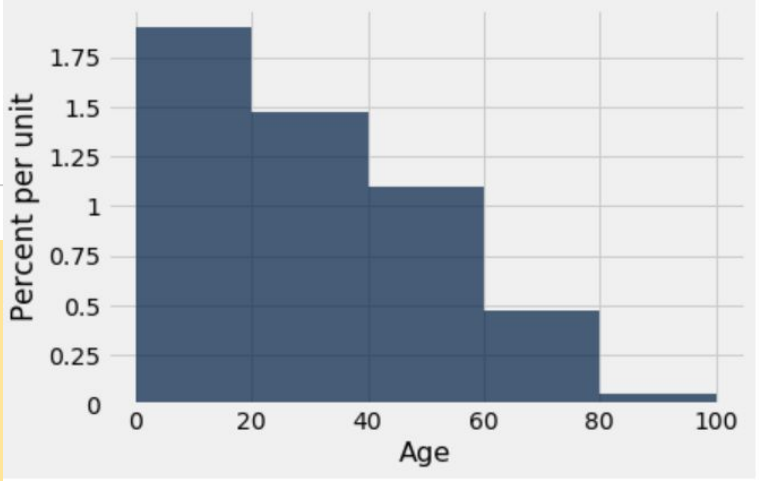


(Demo)

Histogram

- Chart that displays the distribution of numerical values
- Uses bins; one bar for each bin
- Uses the area principle
  - The area of each bar is the percent of individuals in the corresponding bin

Combining Bins



What should happen to our histogram if we combine the two bins [20, 40) and [40, 60) into one large bin [20, 60)?

A. The new histogram should have four bars of equal width.

B. The height of the bar for bin [20, 60) should be the sum of the heights of the bars for bins [20, 40) and [40, 60).

C. The area of the bar for bin [20, 60) should be the sum of the areas of the bars for bins [20, 40) and [40, 60).

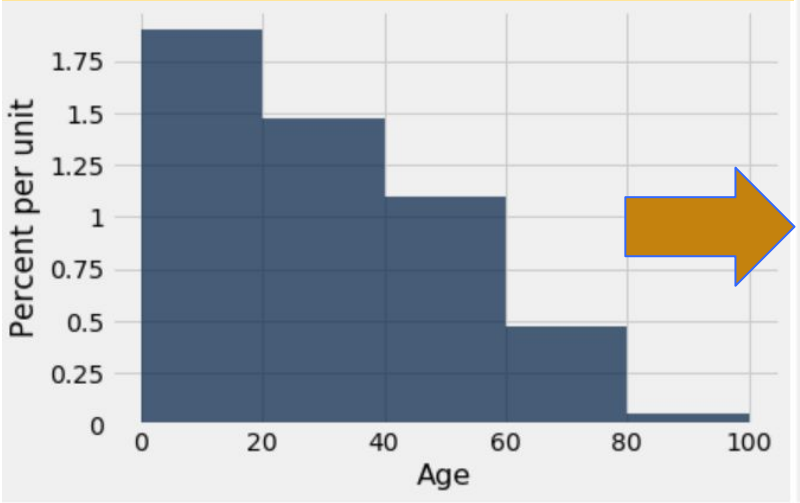
D. More than one of the above.

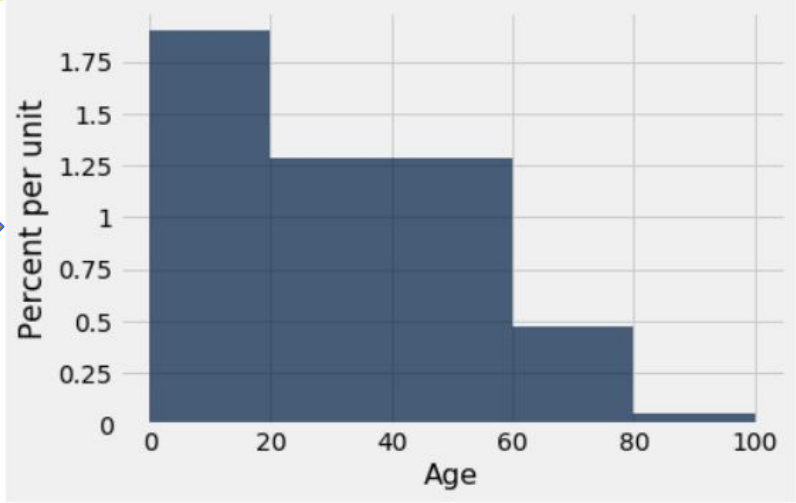
bin	Age count
0	76
20	59
40	44
60	19
80	2
100	0

Combining Bins

What should happen to our histogram if we combine the two bins [20, 40) and [40, 60) into one large bin [20, 60)?

The **area** of the bar for bin [20, 60) should be the sum of the areas of the bars for bins [20, 40) and [40, 60).





bin	Age count
0	76
20	59
40	44
60	19
80	2
100	0

The Density Scale

Histogram Axes

By default, `hist` uses a scale (`normed=True`) that ensures the area of the chart sums to 100%

- The area of a bar is a percentage of the whole
- The horizontal axis is a number line (e.g., years)
- The vertical axis is a rate (e.g., percent per year)

How to Calculate Height

The [0, 20) bin contains 76 out of 200 movies

bin	Age count
0	76
20	59
40	44
60	19
80	2
100	0

How to Calculate Height

The [0, 20) bin contains 76 out of 200 movies

- “76 out of 200” is 38%

bin	Age count
0	76
20	59
40	44
60	19
80	2
100	0

How to Calculate Height

The [0, 20) bin contains 76 out of 200 movies

- “76 out of 200” is 38%
- The bin is 20 - 0 = 20 years wide

bin	Age count
0	76
20	59
40	44
60	19
80	2
100	0

How to Calculate Height

The [0, 20) bin contains 76 out of 200 movies

- “76 out of 200” is 38%
- The bin is 20 - 0 = 20 years wide

$$\text{Height of bar} = \frac{\text{Area}}{\text{Width}}$$

bin	Age count
0	76
20	59
40	44
60	19
80	2
100	0

How to Calculate Height

The [0, 20) bin contains 76 out of 200 movies

- “76 out of 200” is 38%
- The bin is 20 - 0 = 20 years wide

$$\text{Height of bar} = \frac{\text{Area}}{\text{Width}} = \frac{\text{38 percent}}{\text{20 years}}$$

bin	Age count
0	76
20	59
40	44
60	19
80	2
100	0

How to Calculate Height

The [0, 20) bin contains 76 out of 200 movies

- “76 out of 200” is 38%
- The bin is 20 - 0 = 20 years wide

$$\text{Height of bar} = \frac{\text{Area}}{\text{Width}} = \frac{\text{38 percent}}{\text{20 years}}$$

= 1.9 percent per year

bin	Age count
0	76
20	59
40	44
60	19
80	2
100	0

How to Calculate Height

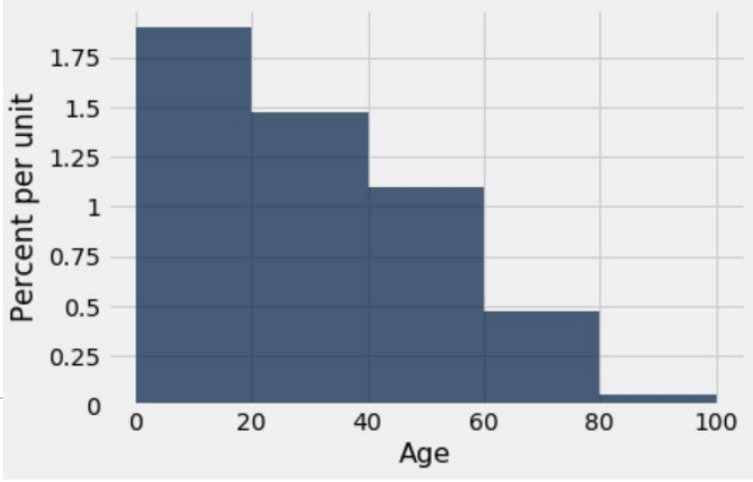
The [0, 20) bin contains 76 out of 200 movies

- “76 out of 200” is 38%
- The bin is 20 - 0 = 20 years wide

$$\text{Height of bar} = \frac{\text{Area}}{\text{Width}} = \frac{38 \text{ percent}}{20 \text{ years}}$$

= 1.9 percent per year

bin	Age count
0	76
20	59
40	44
60	19
80	2
100	0

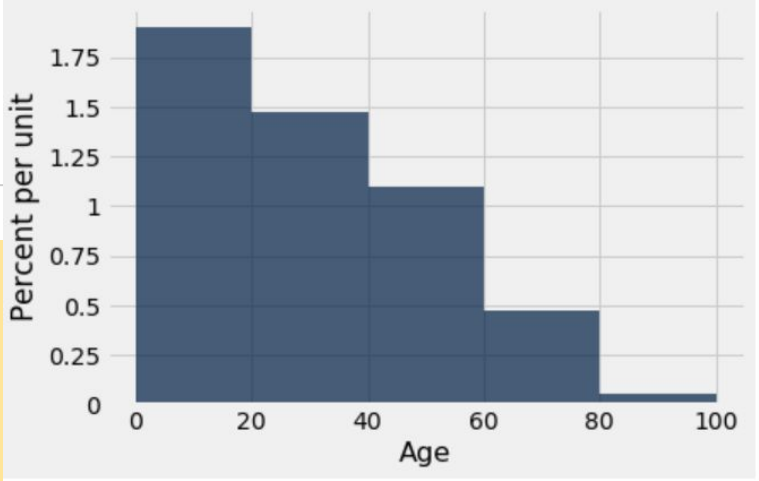


Height Measures Density

$$\text{Height} = \frac{\%\text{ in bin}}{\text{width of bin}}$$

- The height measures the percent of data in the bin ***relative to the amount of space in the bin***.
- So height measures crowdedness, or **density**.

Combining Bins



Suppose we combine the two bins [20, 40) and [40, 60) into one large bin [20, 60). What is the density of the new bin?

A. The new bin has about twice as many movies as each original bin, so it is about twice as dense as each original bin.

B. The new bin is about twice as big as each original bin, so it is about half as dense as each original bin.

C. The new bin has about twice as many movies and is twice as big as each original bin, so it is about the same density as each original bin.

bin	Age count
0	76
20	59
40	44
60	19
80	2
100	0

Height of Combined Bin

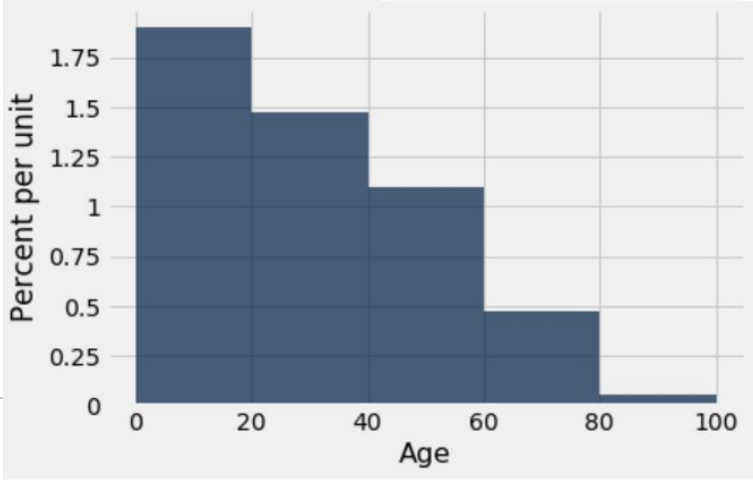
bin	Age count
0	76
20	59
40	44
60	19
80	2
100	0

Combining the [20, 40) and [40, 60) bins creates a [20, 60) bin with 59 + 44 = 103 movies

- “103 out of 200” is 51.5%
- Combined bin is 60 - 20 = 40 years wide

$$\text{Height} = \frac{\text{Area}}{\text{Width}} = \frac{51.5 \text{ percent}}{40 \text{ years}}$$

≈ 1.3 percent per year



Height of Combined Bin

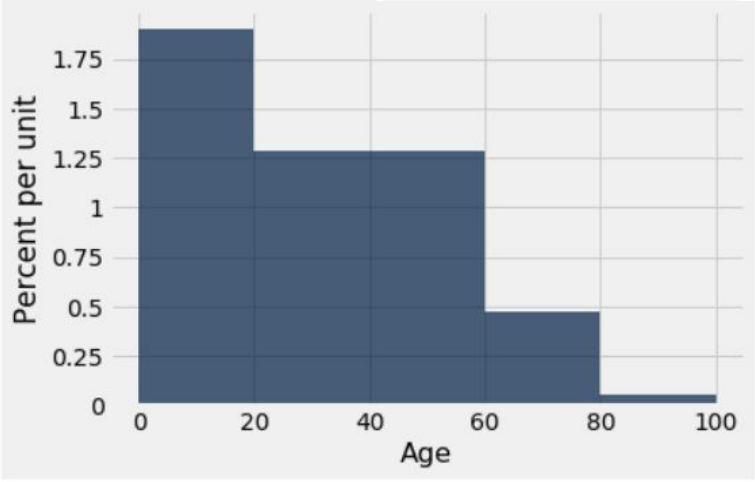
bin	Age count
0	76
20	59
40	44
60	19
80	2
100	0

Combining the [20, 40) and [40, 60) bins creates a [20, 60) bin with 59 + 44 = 103 movies

- “103 out of 200” is 51.5%
- Combined bin is 60 - 20 = 40 years wide

$$\text{Height} = \frac{\text{Area}}{\text{Width}} = \frac{51.5 \text{ percent}}{40 \text{ years}}$$

≈ 1.3 percent per year



Area Measures Percent

Area = % in bin = Height x width of bin

- “How many individuals in the bin?” Use area.
- “How crowded is the bin?” Use height.

(Demo)

Chart Types

Bar Chart vs. Histogram

Bar Chart

- Shows distribution of categorical variable
- Bars have arbitrary (but equal) widths and spacings
- Height (or length) of bar proportional to the percent of individuals

Histogram

- Shows distribution of numerical variable
- Horizontal axis is numerical; to scale with no gaps; can have unequal bins
- Area of bars proportional to the percent of individuals; height measures density

Overlaid Graphs

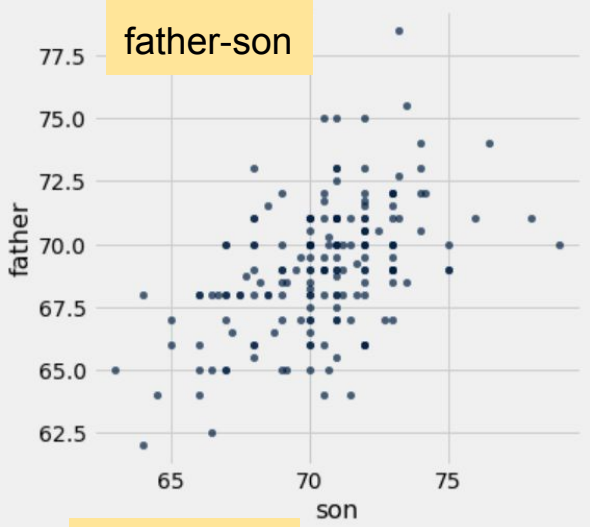
For visually comparing two populations

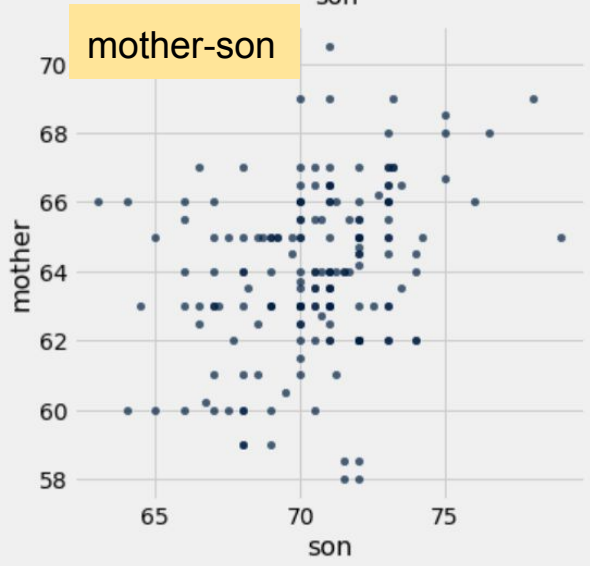
(Demo)

Father or Mother?

Is a son's height more influenced by his father's height or his mother's height?

A. Father, because difference between father and son height is smaller than difference between mother and son height.
B. Mother, because there is more variability in mother's heights than father's heights.
C. Father, because the points on the father-son plot more strongly resemble a line than those on the mother-son plot.
D. Father, because the points on the father-son plot form a steeper curve than the those on the mother-son plot.





Discussion Question

This histogram describes a **year** of daily temperatures in degrees F
(horizontal: temperature (degrees F); vertical: percent per degree F)

Try to answer these questions:

- What proportion of days had a high temp in the range 60-69?
- What proportion had a low of 45 or more?
- What proportion of days had a difference of more than 20 degrees between their high and low temperatures?