DATA 8

Spring 2022

# Lecture 8

Histograms

# Announcements

- Lab 3 is due today at 5pm PT
- HW3 due Thursday, 02/10
    - Turn in on Wednesday for bonus points
- Lab 4 will be released on Monday
- Read this article about causality!

# Weekly Goals

- Monday
  - Table review
  - Working with Census data
- Wednesday
  - Visualizing data
  - Line plots, scatter plots, bar charts
- **Today**
  - Visualizing two kinds of distributions
  - Proportions as areas

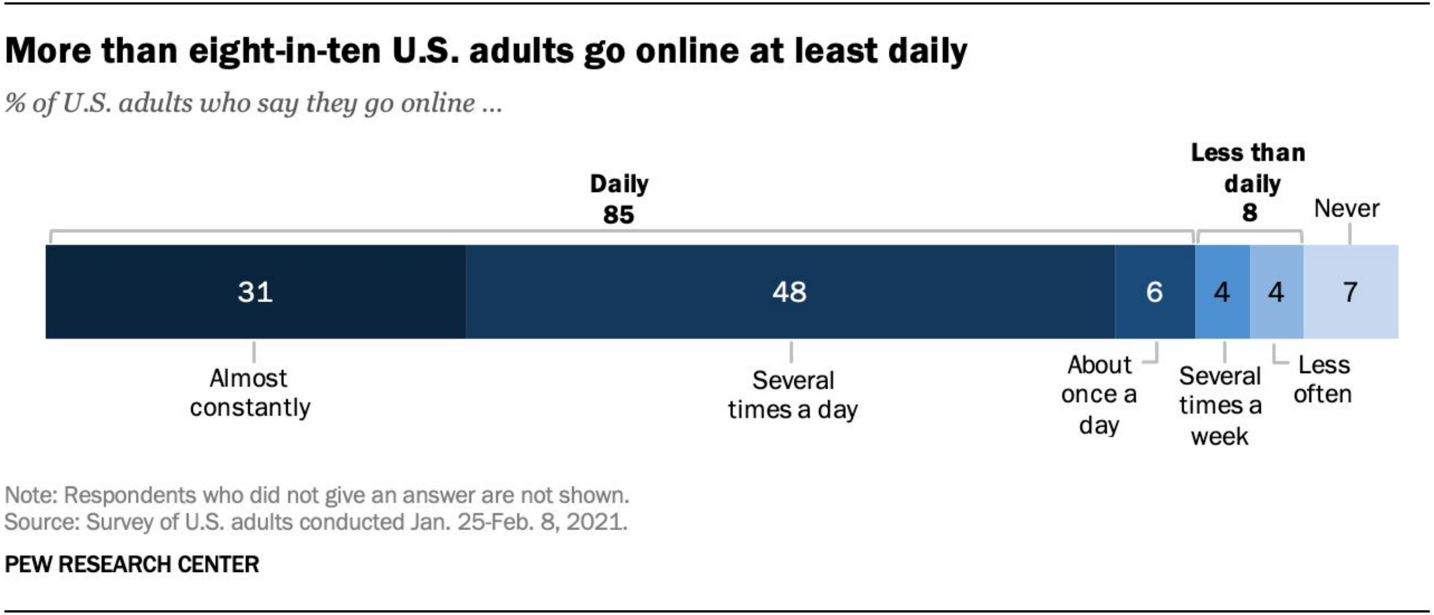
# Distributions

# Terminology

- **Individuals**: those whose attributes are recorded
- **Variable**: an attribute (column)
  - can be **numerical** or **categorical**
  - has different **values**
    - each **individual** has **exactly one value**
  - has a **distribution**:
    - For each different value of the variable, the frequency of individuals that have that value

# A Distribution

Each individual is in exactly one category. Percents add up to 100.



**More than eight-in-ten U.S. adults go online at least daily**

*% of U.S. adults who say they go online …*

Note: Respondents who did not give an answer are not shown.
Source: Survey of U.S. adults conducted Jan. 25-Feb. 8, 2021.

PEW RESEARCH CENTER

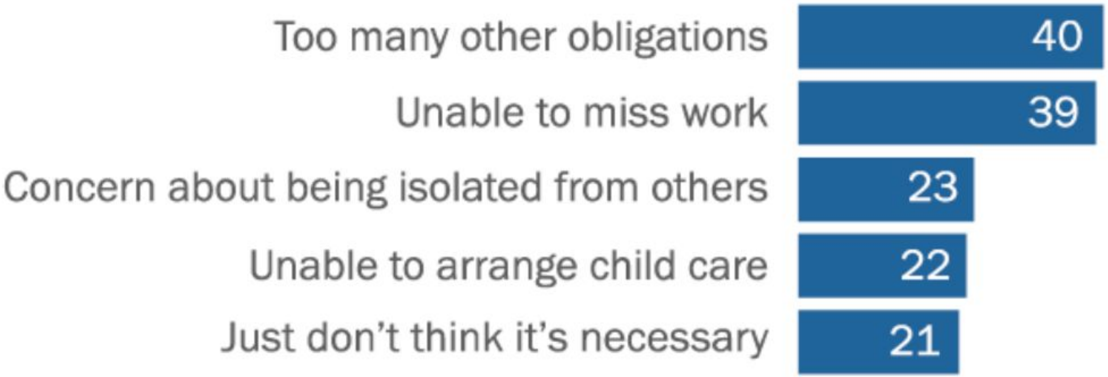
Source: Pew Research

# Not a Distribution

Percents of survey respondents on “a major reason they would find it difficult to quarantine themselves for at least 14 days”

Each respondent can pick more than one answer.

The bars represent overlapping groups.



Source: Pew Research

# Categorical Distributions

(Demo)

# Bar Chart

To display all the values of the variable along with all their frequencies

- Bar chart
  - One bar for each category
  - You can choose the order of the bars
  - Length of bar is the percent (or count) of individuals in that category

(Demo)

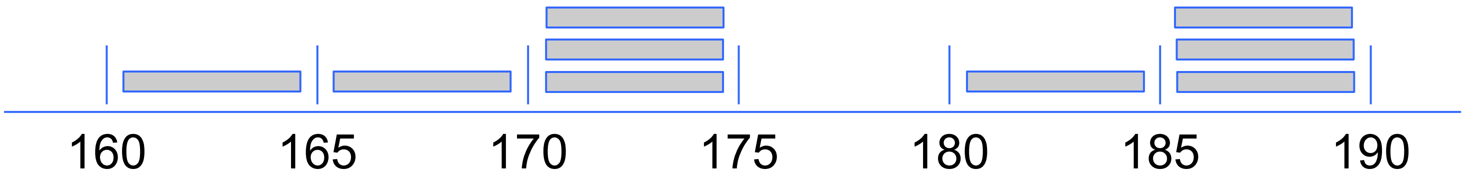
# Numerical Distributions

# Grouping Numerical Values: Binning

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

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

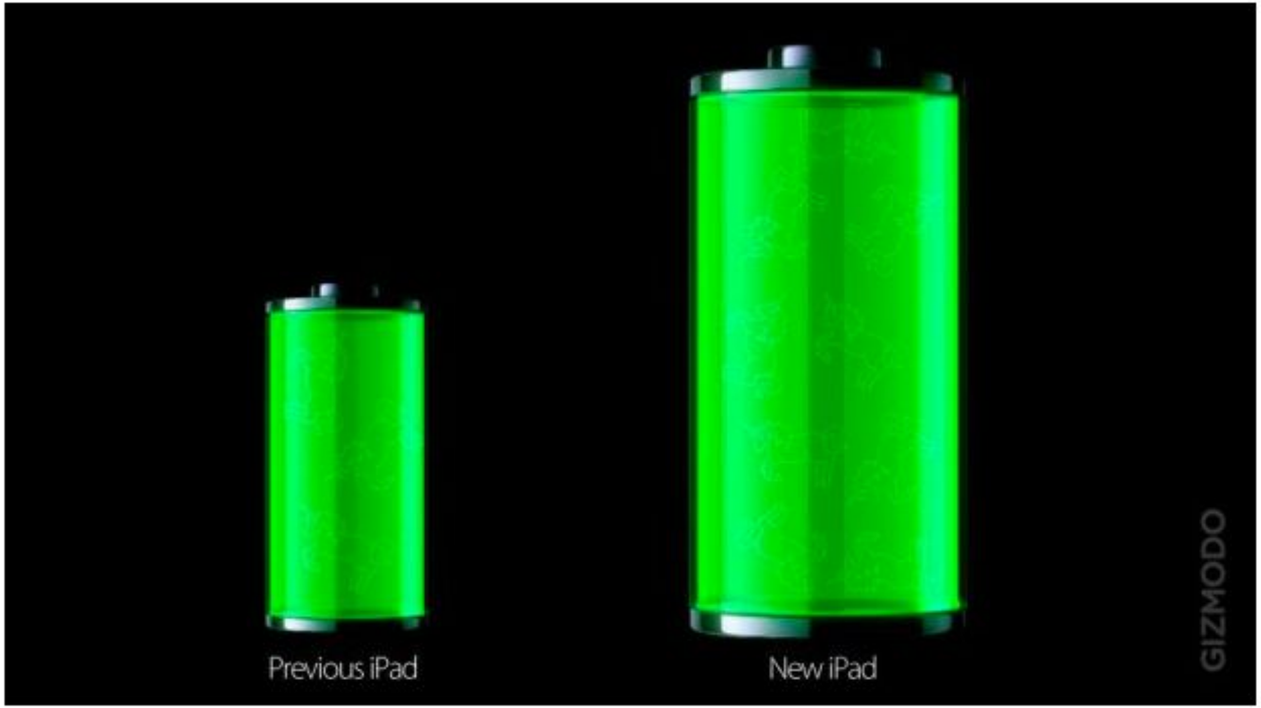


The [185,190) bin

(Demo)

# Area Principle

# What Is Wrong With This Picture?



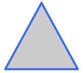
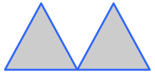
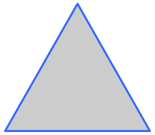
Caption: The new iPad battery is 70% bigger than the previous iPad.

Source

# Area Principle

**Areas** should be proportional to the values they represent.

For example

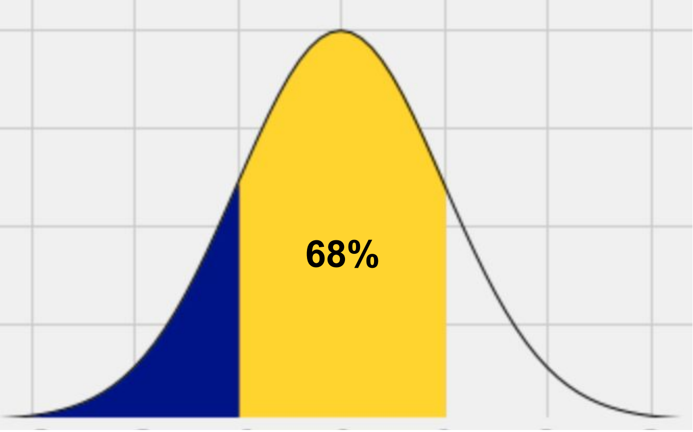
- If you represent 20% of a population by 
- Then 40% can be represented by: 
- But not by: 

# Drawing Histograms

# Histogram

- Displays the distribution of a numerical variable
- One bar corresponding to each bin
- Uses the area principle:
  - The ***area*** of each bar is the ***percent*** of individuals in the corresponding bin

- Later in the course, we will approximate histograms by smooth curves.
  Areas will still represent percents.



(Demo)

# Density

# Histogram Axes

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

- The **area** of each bar is a percentage of the whole

- The horizontal axis is a number line (e.g. years), and the bins sizes don’t have to be equal to each other

- The vertical axis is a rate (e.g., percent per year)

(Demo)

# How to Calculate Height

The [40, 65) bin contains 56 out of 200 movies

- “56 out of 200” is 28%
- The bin is 65 - 40 = 25 years wide

$$\text{Height of bar} = \frac{\text{28 percent}}{\text{25 years}}$$

$$= \text{1.12 percent per year}$$

# Height Measures Density

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

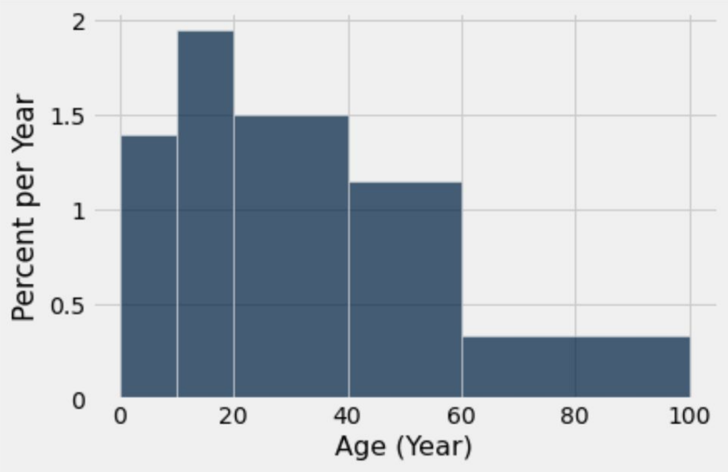
- The height measures the percent of data in the bin ***relative to the amount of space in the bin*****.**
- Height measures crowdedness, or **density**.
- Units: percent per unit on the horizontal axis

# Area Measures Percent

**Area of bar = % in bin = Height x width of bin**

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

# Discussion Questions



Compare the bins [10, 20) and [20, 40).

- Which one has more movies?
  **Answer: [20, 40), bigger area**

- Which one is more crowded?
  **Answer: [10, 20), taller**

# Bar Chart or Histogram?

To display a distribution:

## Bar Chart

- Distribution of categorical variable
- Bars have arbitrary (but equal) widths and spacings; in any order
- **height (or length)** and **area** of bars proportional to the percent of individuals

## Histogram

- Distribution of numerical variable
- Horizontal axis is numerical: drawn to scale, no gaps, bins can be unequal
- **Area** of bars proportional to the percent of individuals; **height** measures density

# Discussion Questions

What is the height of each bar in this histogram?

```
my_bins = make_array(0, 15, 25, 85)
incomes.hist(1, bins = my_bins)
```

What are the vertical axis units?

| Name | 2016 Income (millions) |
|---|---:|
| Jennifer Lawrence | 61.7 |
| Scarlett Johansson | 57.5 |
| Angelina Jolie | 40 |
| Jennifer Aniston | 24.75 |
| Anne Hathaway | 24 |
| Melissa McCarthy | 24 |
| Bingbing Fan | 20 |
| Sandra Bullock | 20 |
| Cara Delevingne | 15 |
| Reese Witherspoon | 15 |
| Amy Adams | 15 |
| Kristen Stewart | 12 |
| Amanda Seyfried | 10.5 |
| Tina Fey | 10.5 |
| Julia Roberts | 10 |
| Emma Stone | 10 |
| Natalie Portman | 8.5 |
| Margot Robbie | 8 |
| Meryl Streep | 6 |
| Mila Kunis | 4.5 |

# Answers

Vertical axis units: Percent per million

```
my_bins = make_array(0,15,25,85)
[0, 15): (45%)/(15 million)
         = 3 % per million
[15, 25): (40%)/(10 million)
         = 4 % per million
[25, 85): (15%)/(60 million)
         = 0.25 % per million
```

| Name | 2016 Income (millions) |
|---|---:|
| Jennifer Lawrence | 61.7 |
| Scarlett Johansson | 57.5 |
| Angelina Jolie | 40 |
| Jennifer Aniston | 24.75 |
| Anne Hathaway | 24 |
| Melissa McCarthy | 24 |
| Bingbing Fan | 20 |
| Sandra Bullock | 20 |
| Cara Delevingne | 15 |
| Reese Witherspoon | 15 |
| Amy Adams | 15 |
| Kristen Stewart | 12 |
| Amanda Seyfried | 10.5 |
| Tina Fey | 10.5 |
| Julia Roberts | 10 |
| Emma Stone | 10 |
| Natalie Portman | 8.5 |
| Margot Robbie | 8 |
| Meryl Streep | 6 |
| Mila Kunis | 4.5 |