DSFA

Spring 2021

Lecture 8

Histograms

Announcements

- Homework 3 out today, due next Friday 3/5.

When poll is active, respond at **PollEv.com/dsfa**

Let's check if you are registered for PollEverywhere!

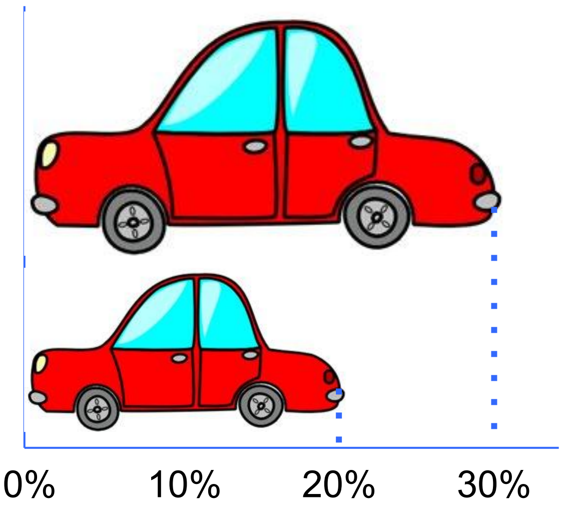
I am registered!

I'm not registered.

Start the presentation to see live content. For screen share software, share the entire screen. Get help at **pollev.com/app**

Area Principle

Areas should be proportional to the values they represent

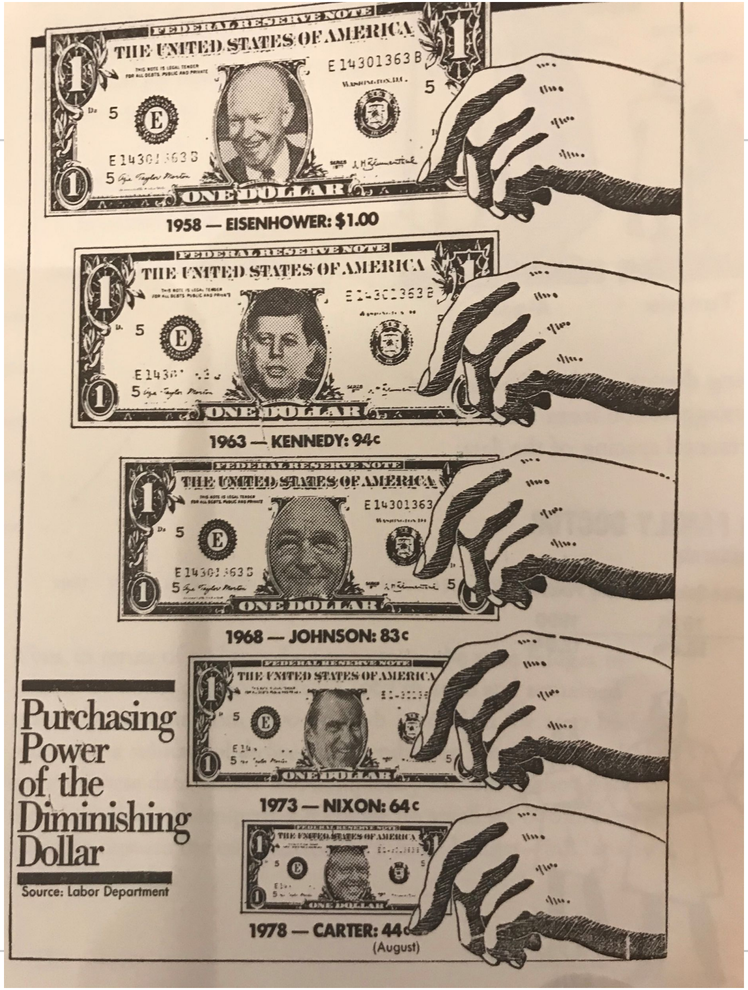


In 2013,

30% of accidental deaths of males were due to automobile accidents

20% of accidental deaths of females were due to automobile accidents

Example from Tian Zheng



From Tufte, p. 70:
“If the area of the dollar is accurately to reflect its purchasing power, then the 1978 dollar should be about twice as big as that shown.”

Bar Charts (Review)

Types of Data

All values in a column should be both the same type **and** be comparable to each other in some way

- **Numerical** — Each value is from a numerical scale
    - Numerical measurements are ordered
    - Differences are meaningful
- **Categorical** — Each value is from a fixed inventory
    - May or may not have an ordering
    - Categories are the same or different

Bar Charts of Counts

Distributions:

- The distribution of a variable (a column) describes the frequency of its different values
- The `group` method counts the number of rows for each value in a column

Bar charts can display the distribution of categorical values

- Proportion of how many US residents are male or female
- Count of how many top movies were released by each studio

(Demo)

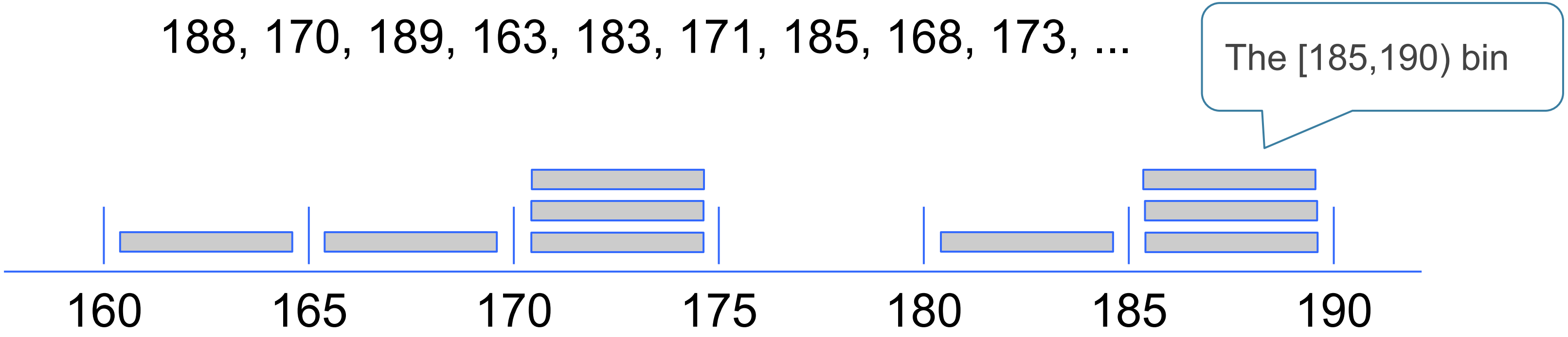
Binning

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

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

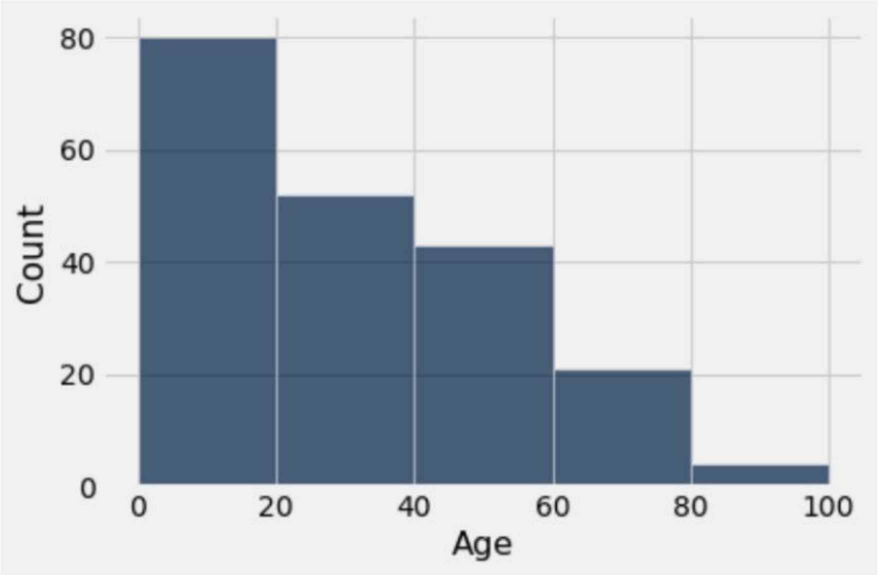


Histogram

Chart to display the distribution of numerical values using bins

(Demo)

For this histogram

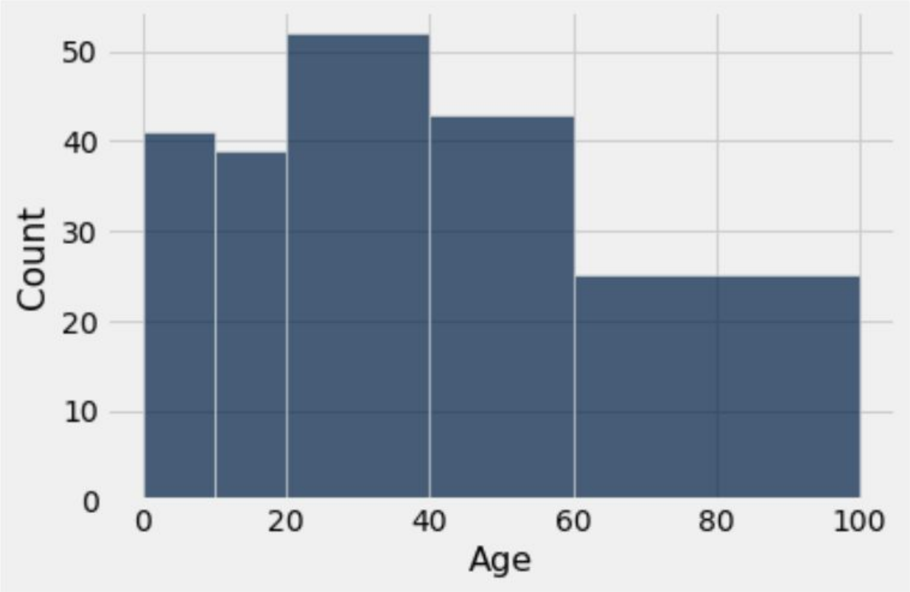


The heights are proportional to the counts in each bin

The areas are proportional to the counts in each bin

Both heights and areas are proportional to the counts in each bin

For this histogram



Heights are proportional to the counts in each bin

Areas are proportional to the counts in each bin

Both heights and areas are proportional to the counts in each bin

The Density Scale

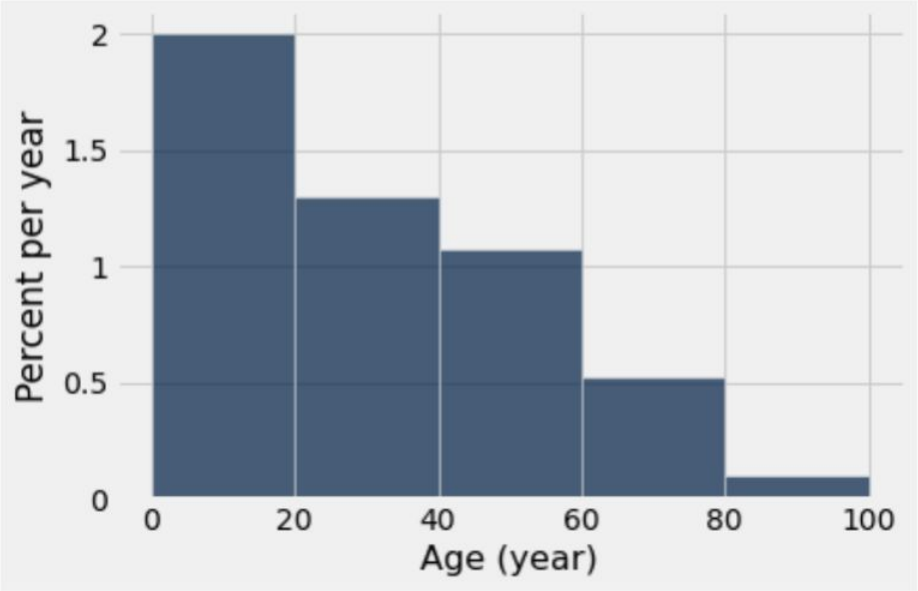
Histogram Axes

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

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

(Demo)

For this histogram



Heights are proportional to counts in each bin

Areas are proportional to counts in each bin

Both heights and areas are proportional to counts in each bin

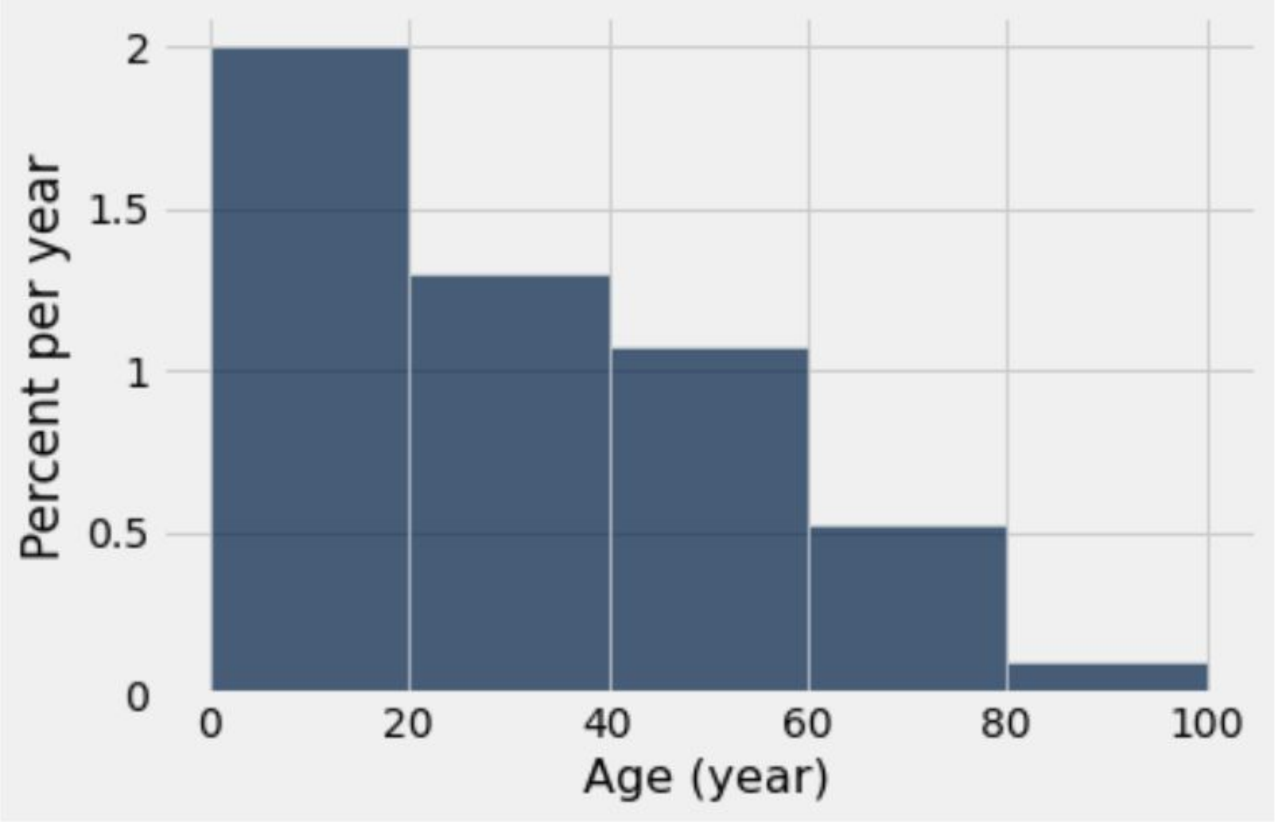
How to Calculate Height

The [0, 20) bin contains 80 out of 200 movies

- “80 out of 200” is 40%
- The bin is 20 - 0 = 20 years wide

$$\text{Height of bar} = \frac{40 \text{ percent}}{20 \text{ years}}$$

$$= 2 \text{ percent per year}$$


2
1.5
1
0.5
0
Percent per year
0
20
40
60
80
100
Age (year)

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**.

(Demo)

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.

Discussion Question

What's the height of each bar in these two histograms?

```
actress.hist(1, bins=[0,15,25,85])
```

```
actress.hist(1, bins=[0,15,35,85])
```

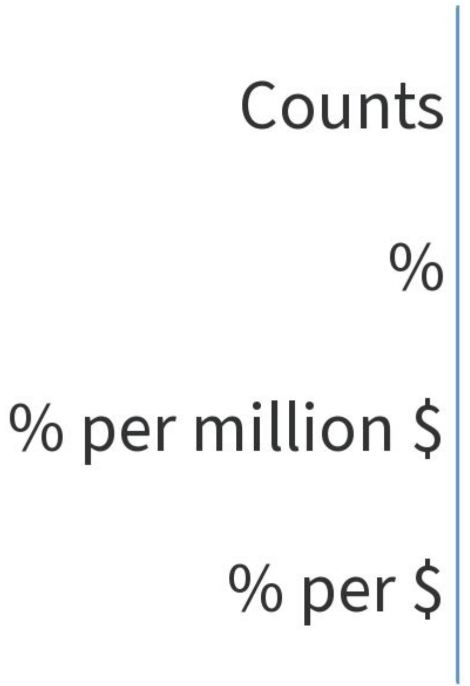
What are the vertical axis units?

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

When poll is active, respond at **PollEv.com/dsfa**

Text **DSFA** to **22333** once to join

What would be the units of the vertical axis?



Counts

%

% per million $

% per $

Chart Types

Bar Chart Versus Histogram

Bar Chart

- 1 categorical axis &
  1 numerical axis
- Bars have arbitrary (but equal) widths and spacings
- For distributions: height (or length) of bars are proportional to the percent of individuals

Histogram

- Horizontal axis is numerical, hence to scale with no gaps
- Height measures density; areas are proportional to the percent of individuals

Overlaid Graphs

For visually comparing two populations

(Demo)