DATA 8

Summer 2018

Lecture 7

Histograms

Slides created by John DeNero (denero@berkeley.edu) and Ani Adhikari (adhikari@berkeley.edu)
Contributions by Fahad Kamran (fhdkmrn@berkeley.edu) and Vinitra Swamy (vinitra@berkeley.edu)

Announcements

Types of Data

Two Important Types

- **Numerical** — Each value is from a numerical scale
  - Ordered, because they are numbers
  - Differences, averages, etc are meaningful

- **Categorical** — Each value is from a fixed inventory
  - May or may not have an ordering

Terminology

- **Individuals**: those whose features are recorded
- **Variable**: a feature, an attribute
- A variable has different **values**
- Values can be **numerical** or **categorical**, and of many sub-types within these
- Each **individual has exactly one value** of the variable

- **Distribution**: For each different value of the variable, the frequency of individuals that have that value

Categorical Distributions

Visualization

- Bar charts are commonly used to visualize categorical distributions
- One axis is categorical, one numerical

(Demo)

Displaying a Categorical Distribution

- The distribution of a variable (a column, e.g. Studios) describes the frequencies of its different values
- The `group` method counts the number of rows for each value in the column (e.g. the number of top movies released by each studio)
- Bar charts can display the distribution of a categorical variable (e.g. studios):
  - One bar for each category
  - Length of bar is the count of individuals in that category

Binning a Numerical Variable

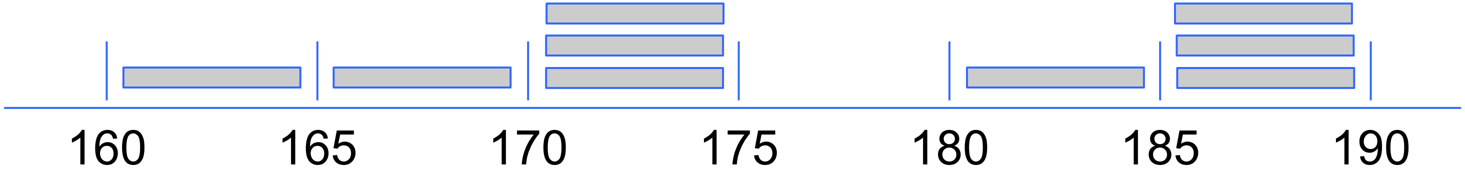
Binning Numerical Values

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, ...

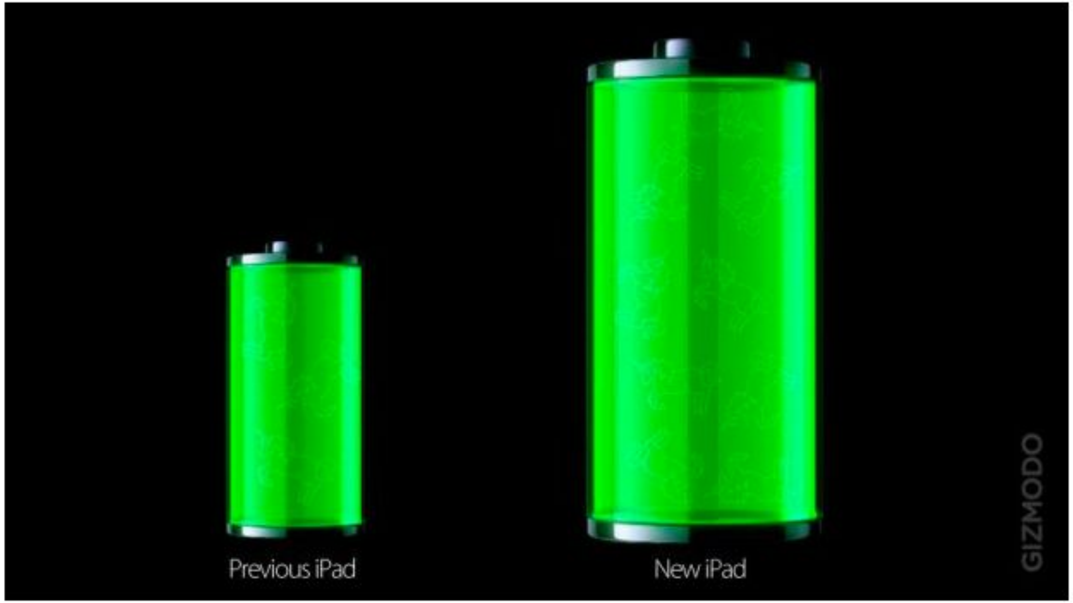
The [185,190) bin

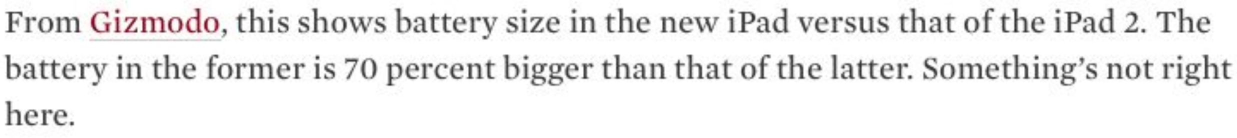


(Demo)

Area Principle

What Is Wrong With This Picture?



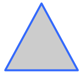

From Gizmodo, this shows battery size in the new iPad versus that of the iPad 2. The battery in the former is 70 percent bigger than that of the latter. Something's not right here.

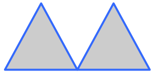
From Flowing Data:
http://flowingdata.com/2012/03/16
/new-ipad-battery-size-is-huge/

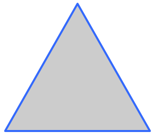
Area Principle

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

- Chart that displays the distribution of a numerical variable

- Uses bins; there is one bar corresponding to each bin

- Uses the area principle:
    - The ***area*** of each bar is the percent of individuals in the corresponding bin

(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 52 out of 200 movies

- “52 out of 200” is 26%
- The bin is 65 - 40 = 25 years wide

$$\text{Height of bar} = \frac{\text{26 percent}}{\text{25 years}}$$

$$= \text{1.04 percent per year}$$

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.

Disclaimer about Histograms

- Height measures crowdedness in a bin, but it does **not** tell you how the data is distributed within a bin

- Imagine a bin 20-25 containing 20% of the data
  - What can we say about the number of values in the interval 20-21?
  - What about 21-22?

Bar Chart or Histogram?

To display a distribution:

Bar Chart

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

Histogram

- Distribution of numerical variable
- Horizontal axis is numerical: 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?

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

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

Answers

Vertical axis units: Percent per million

```
incomes.hist(1, bins=[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